MetaScholar Initiative
Emory University

General Libraries
540 Asbury Circle
Emory University
Atlanta, GA 30322

Phone 404 727 2204
Fax 404 727 0827

MetaArchive Project Final Report

A project for the development and evaluation of an Open Source OAI-compliant infrastructure for surfacing hidden scholarly resources

October 2001 – October 2003

Last Revised Monday 01-March-2004

Table of Contents

Executive Summary

The MetaArchive project sought to demonstrate the applicability of a cross-institutional scholarly service for metadata harvesting, searching, and subject organization. The project developed and evaluated open source software tools for creating infrastructures for surfacing hidden scholarly resources based on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). A process model was investigated whereby a centralized project team based at a large research institution could facilitate smaller archives in disseminating metadata about their collections.

The following are highlights of project results:

- Several open source tools including the Open Digital Libraries software from Virginia Tech and the ARC software from Old Dominion University were used to produce an effective, scalable, and flexible infrastructure for metadata providers and harvesting operations.
- 23 OAI data providers were created at 16 libraries and archives in the course of the project, serving out 62,019 records, associated with some 420 different collections.
- The problem of metadata context collisions was examined in the resulting metadata, providing insight into the entire cycle of metadata production and consumption, or metadata gardening.
- Potential audiences for the service were studied, with a series of focus groups and discussions providing feedback concerning scholarly portal functions that might be developed in conjunction with metadata harvesting operations.
- A sustainability study was undertaken using perspectives gained from the project to model patterns in how scholarly portal operations based on metadata harvesting might logically progress in scale and be supported over time.

The project has produced a body of expertise concerning metadata harvesting using the OAI-PMH and associated open source software tools for such operations. The project has also identified a number of areas for future research.

1. Review of Problems and Opportunities Addressed

The MetaArchive project was a research and feasibility study of the opportunities provided for surfacing hidden scholarly collections by means of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). It was undertaken as one of a series of projects sponsored by the Andrew W. Mellon Foundation to advance the understanding and practice of the OAI-PMH for scholarly communication purposes.

Archives and other repositories of scholarly information face many difficulties in exposing information about their collections to potential users. This issue has received much attention in recent years, and concern over the issue of exposing relatively "hidden" collections has surfaced in the form of various workshops and white papers [Jones, 2003].

The OAI-PMH offers a standard mechanism for flexibly disseminating metadata from repositories for discovery purposes. The MetaArchive project sought to investigate some specific applications of the protocol for sharing metadata between archives; hence the name, *MetaArchive.*

1.1 Summary Points from Original Project Proposal

> The unique resources maintained in academic libraries often lack the needed visibility to reach the scholars who would be interested in such materials. New standards for information exchange enable new means of scholarly information discovery. Emory University proposes a Metadata harvesting and searching service that will offer multiple institutions a combination of new technologies for sharing information about locally maintained resources of interest to scholars, as well as a means of seeking and discovering complementary information held by other institutions.
>
> This service will aggregate metadata (information about scholarly information) from contributing institutions, and will provide a publicly searchable web interface to this metadata aggregation. The web interface will enable both comprehensive searches and searching strategies targeted to particular subject domains.

1.2 Additional Background Comments

The project was based at the General Libraries of Emory University in Atlanta, Georgia. It focused on assembling metadata concerning two subject domains held by small archives in a group of collaborating Southern libraries: 1) archived papers of major political figures, and 2) religious institutional records. Martin Halbert, Director for Library Systems at Emory University, was the principal investigator of the project, and is the primary author of this report.

As the MetaArchive project began, the SOLINET non-profit corporation (also based in Atlanta) also received a grant to undertake another Mellon metadata harvesting project called AmericanSouth. SOLINET subsequently approached Emory University with the idea that the AmericanSouth project might be conjoined with MetaArchive, with all AmericanSouth activities taking place at Emory. This was done, and the resulting conjunction of projects was termed the MetaScholar Initiative.

2. Infrastructure Outcomes

The MetaArchive project set out to answer several questions through infrastructure goals. The advent of the OAI-PMH in 2001 appeared to offer many new opportunities for disseminating metadata from repositories of cultural information for scholarly communication purposes, but only if the protocol could be easily deployed through readily available systems that required modest levels of investment by such repositories.

Could OAI provider and harvester capabilities be established at research libraries using open source tools, with modest levels of development efforts? If so, then the protocol could conceivably open the door to an entire new generation of collaboratively developed discovery systems for researchers. However, these opportunities would be curtailed if the protocol proved difficult to implement on top of existing digital library infrastructures, could only be implemented with expensive commercial tools requiring large capital investments, could not easily lead to searchable union databases, or did not scale well. The MetaArchive project therefore set forth several goals to make practical advances in answering these questions.

Outcomes set forth in original project proposal:

> Emory will create an infrastructure capable of:
>
> a. Responding to OAI protocol requests for metadata feeds.
>
> b. Harvesting metadata from other OAI-compliant servers.
>
> c. Providing a public web interface able to perform keyword and field delimited searches on the metadata records aggregated in the system.
>
> d. The web interface will also have some portal functions aimed at particular scholarly communities organized around subject domains. Which communities and subject domains will be cultivated in this way will be determined in the course of the project based on feedback from scholars, logical strengths and complimentary features of metadata harvested from contributing institutions.

2.1 OAI Data Providers

We sought to identify a core set of open source software tools with which we could implement OAI data providers. Two main tool sets were used: the ODL software, and ad hoc data conversion and formatting scripts written in PERL and XSLTproc. Open source was clearly the best approach to creating these OAI data providers; a small set of tools provide all the functionality we needed to add OAI capabilities to a wide variety of digital library infrastructures.

2.1.1 Open Digital Libraries (ODL) Software

The primary open source software used to create OAI data provider systems was the Open Digital Libraries (or ODL) software suite developed by Hussein Suleman at the Digital Library Research Laboratory at Virginia Tech University in 2002. The ODL software was created as an attempt to develop a set of flexible digital library (or DL) components for recombinant construction of DL

infrastructures. The ODL software is fundamentally based on the OAI-PMH as a unifying mechanism for interoperation, and includes a set of flexible OAI data provider modules (see http://www.dlib.vt.edu/projects/OAI/#software).

The primary ODL modules that we used to create OAI providers were the *ODL XML File-based OAI Data Provider*, and the *OAI-PMH2* version of this module (when it became available after the release of version 2.0 of the protocol). These modules were written in PERL, and had many features that made them adaptable to a variety of situations. For example, during the project we often set up the ODL software on a file directory structure of static XML, SGML, HTML records, in which case the ODL software would respond to OAI queries by serving out XML Dublin Core representations of the underlying data. The Auburn providers are a good example of this scenario. In other cases, such as the records in Emory's Nunn digital archive, the ODL software was adapted to provide XML output from existing DBMS systems.

Full details concerning the total numbers of providers and characteristics of the resulting metadata are provided in section 3. What is worth noting here is the degree to which a single open source tool was able to be adapted to a large number of situations. Of the 23 provider infrastructures created in the course of the project, 16 of them (or roughly 70%) were implementations of the ODL software, resulting in exposure of some 8,162 records. Nor was the ODL software solely reimplemented by MetaScholar staff; a number of digital library staff at other institutions were readily able to deploy the software. Examples of this situation include the UVA and UNC providers that we provided consultative advice for, but ultimately did not implement.

The flexibility of the ODL software is a tribute to its developer, Dr. Hussein Suleman, an active force in the OAI implementers group and now a professor at the University of Capetown. Suleman was one of the consultants we worked with through the AmericanSouth project, and he was a great advisor and information resource concerning many aspects of the OAI-PMH.

2.1.2 Other Provider Tools

In 2 cases we developed ad hoc providers from scratch, using a combination of PERL and XSL transforms. Our programming team was able to develop these providers relatively easily, and one of these providers serves out one of the large blocks of metadata developed in the project, the 18,186 records of Southwestern University.

Clearly, OAI data providers based on readily available open source software tools can easily be deployed on top of existing digital library infrastructures, provided that programmers with some understanding of the protocol and basic programming techniques are available.

Our experience indicated that a small team of such programmers could easily catalyze the establishment of a large number of OAI data providers. The main impediments to wider deployment of the OAI-PMH is not technical, they are organizational. This topic will be taken up in section 3 of this report.

2.2 Metadata Harvesting Systems

A variety of technical solutions for metadata harvesting were investigated in the course of the first year of the project, as documented separately in the MetaArchive interim project report. The ultimate tool selected was the ARC software developed at Old Dominion University.

2.2.1 ARC Software

The ARC software was developed at Old Dominion as a testbed service for OAI harvesting [Maly, 2003]. The ARC service was the first major attempt to comprehensively harvest all extant OAI providers, and has subsequently scaled well to aggregations of millions of records.

Early in the project, the MetaArchive project staff became aware of the ARC software in the course of canvassing the field for metadata harvesting tools. At that point the ARC software was a proprietary research tool of Old Dominion University. It was therefore not being considered as a component to be used in the project, since our focus was solely on open source tools. A subsequent chance conversation between Dr. Kurt Maly and Martin Halbert indicated that Old Dominion might be willing to make the software available under an open source license.

The MetaScholar Initiative subsequently contacted the Old Dominion digital library research group to further discuss this possibility. After discussions of the advantages to an open source deployment, the researchers at Old Dominion became convinced that this course of action was highly desirable and took action to release the software through SourceForge.

We are very pleased that the MetaScholar Initiative was instrumental in urging the Old Dominion group to release the ARC software as open source, and feel that this was a significant case of open source advocacy by the project.

The ARC software code was downloaded, and redeployed at Emory University, in consultation with research staff members at Old Dominion, who were very helpful in taking time to explain the operation of the software, which was almost completely undocumented. The project subsequently donated notes on implementing the software to Old Dominion for possible inclusion in future releases of the software.

Tests of the ARC software showed that it was indeed robust. As a load test, we harvested and indexed the entire holdings of the Library of Congress American Memory OAI data provider. The more than 100K records in this provider gave the system no problems.

ARC provided a comprehensive solution to both harvesting, indexing and searching. The only real problem was that ARC was entirely written in Java, with relatively little ability to interface with other systems. We were interested in experimenting with various portal technologies, so we wanted to develop some mechanism for bridging the ARC system with other open source tools.

2.2.2 Bridge Systems

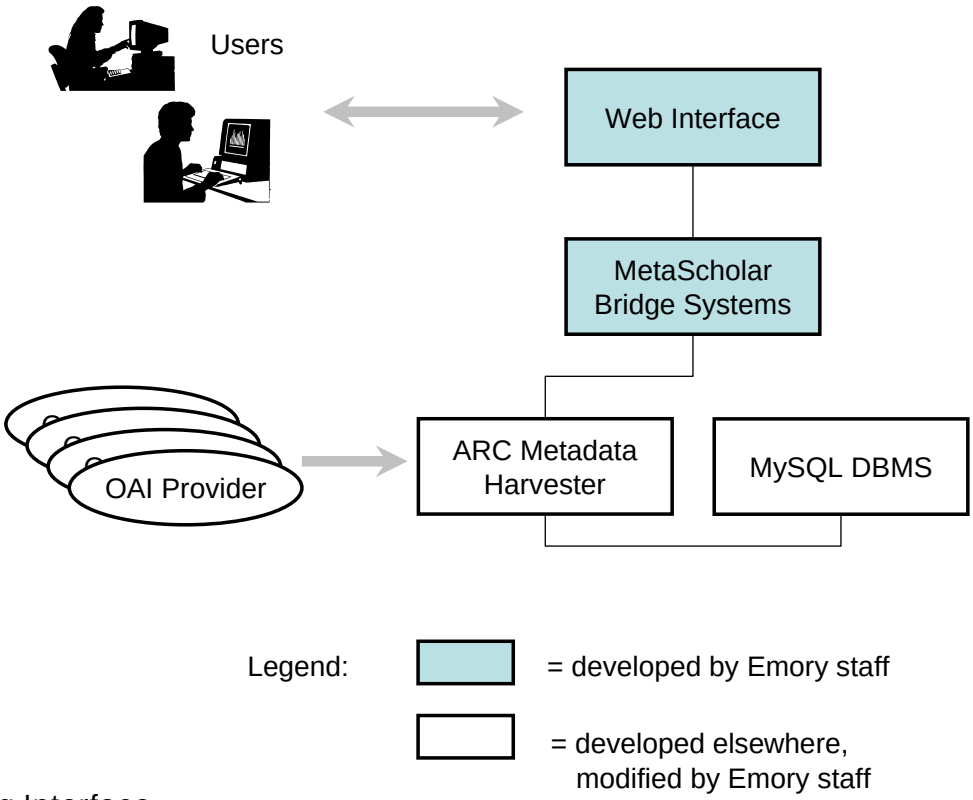
We set out to develop a set of bridge software tools for customization of the ARC interface, and in the process created an interoperable application programming interface for the ARC software. This provided a capability for integrating ARC with a variety of other systems.

The bridge systems were written in PERL, and provided a number of generic interfaces that enabled various kinds of connections with other software. We experimented with a number of open source content management and portal environments, including Zope and OpenCMS, before selecting the PostNuke software for further work.

At roughly this point in the project, the decision was made to combine the portal that we envisioned creating for MetaArchive with the portal we were simultaneously developing for the AmericanSouth project. Because of the overlap in subject domains between the projects, it did not make sense to developing separate portals, and made a great deal of sense to merge the systems conceptually to pool more metadata. From this point forward, the MetaArchive team worked primarily on portal development in the AmericanSouth context.

The bridge systems we developed were an essential step in continued development of ARC's capabilities. These bridge systems are being made available as open source software to interested parties (NCSU has recently expressed interest in this software).

Figure 1. MetaArchive Technical Infrastructure



2.3 Searching Interface

The ARC software provided an excellent search interface, so development efforts in this area could be devoted to understanding the needs of scholars seeking to search aggregated union databases of harvested metadata. Our results in this area led to many findings about what functions and approaches that work well and those that do not work so well. Since our search interface results are detailed at some length in the companion final report of the AmericanSouth project, they will not be repeated here.

The key point to make about searching interfaces for harvested metadata is that they must be simple to use, but must produce sophisticated results that users can easily understand. OAI-PMH harvesting systems make it possible to quickly aggregate a large amount of heterogeneous metadata through automated means. A surprising finding of our project is that searching harvested metadata is not enough by itself, users strongly desire systems that can simultaneously search both harvested metadata and other kinds of scholarly content relevant for research in particular subject domains.

2.4 Portal Functions

The project sought particularly to investigate the following questions, "Can metadata harvesting form the foundation for a portal system of benefit to scholarly communication? If so, what are the characteristics of such a scholarly portal?"

The concept of scholarly portals had received much attention in research libraries in the months leading up to 2001. Jerry Campbell had published a provocative white paper articulating the case for a scholarly portal to web resources [Campbell, 2000], a paper that ultimately led to the establishment of the ARL Scholars Portal Working Group effort.

Investigating the promise of the OAI-PMH for scholarly portal functions was a significant aspect of the original challenge of the Mellon Foundation to those engaging in the planning that ultimately led to the Mellon metadata harvesting projects. The MetaArchive project sought to carefully consider this question by engaging a Subject Portal Working Group of librarians and archivists in experiments to try out some speculative concepts of OAI-based scholar's portals. While the AmericanSouth project is the ultimate product of these experiments, we here note some more general conclusions about this issue.

For the print universe, the traditional library is the scholar's portal to knowledge. The traditional library collects and organizes access to a tremendous range of printed expressions of information. It customizes the experience of these resources by scholars through personal interaction with librarians and other facilitators.

For the electronic universe, the digital library is the equivalent scholar's portal to knowledge. In the discussions of the working group formed at Emory to examine this issue, the following general points were extrapolated from this equivalency, with the practical results noted:

- The digital library must provide organized access mechanisms for researchers who are engaging in both known item searching and unknown item browsing. While keyword searching provides a means for identifying known items in collections of harvested metadata, browsing requires more structure across collections of metadata in terms of paths like subject headings, dates, and creators. Because of the issue of metadata context collisions (described later in this report, such browsing functions are essentially disabled in aggregations of harvested metadata that are currently possible. As this issue became clear, we began to explore conceptual means of structuring and remediating metadata post-harvest. This ultimately led to the MetaCombine project proposal, an attempt o develop technologies for clustering metadata in various ways, and thereby providing the consistency of structure currently unavailable in metadata that can be harvested.
- Like traditional libraries, digital libraries must have coherent collection development boundaries defined by the needs of its clientele in order to be a useful lens on the universe of information. The point here is that ongoing collection development is a significant work activity for any digital library. The initial collection development effort that went into MetaArchive and AmericanSouth consisted solely of identifying metadata from a modest number of complementary collections. To become an effective digital library, ongoing collection development efforts are needed to solicit feedback from users and systematically cultivate the collection. The lack of attention to this ongoing function is part of the reason for the unevenness of the resulting collections in both AmericanSouth and MetaArchive. We now think that a comprehensive collection scanning activity will be needed to identify and address gaps in the virtual collection we have assembled on southern culture and history. Attention must be given to involving sufficient experts to act as an effective collection development group. This group should probably include a mix of professors, librarians, and archivists. This will be the focus of a future project in the MetaScholar Initiative.

- Unlike traditional libraries (at least since the monastic era), digital libraries are today actively engaged in the production of new knowledge, and wholesale transcription efforts from one medium (print) to another (digital). This activity must be factored into the cultivation of the digital library as a resource for particular scholarly groups, with strategic investments in not only digitization projects but especially also partnerships with scholars to create knowledge in the new medium. Scholars welcome this partnership, as they see many advantages of work with the combination of technologists and librarians represented in digital library projects. In forming the advisory board of scholars and other experts for the MetaCombine project, we are closely examining the opportunities for such new forms of knowledge creation based on many of the ideas for new projects that emerged from the discussions in MetaArchive and AmericanSouth. One of the first proposals from the MetaCombine advisory board in this regard is a new peer-reviewed online forum to be called Southern Spaces, which will build on the new concept of regionality and spatial conceptions of culture that the advisory board has focused on.

3. Metadata Aggregation Outcomes

Perhaps the key goal of the MetaArchive project was to develop and test the feasibility of a model of collaborative metadata aggregation in which a central project staff provided technical expertise and services to a cluster of smaller institutions. What was being assessed was whether the OAI-PMH would in fact prove to be an effective structuring process for pooling metadata originating in smaller institutions, especially for archives that would otherwise be unlikely to expose such metadata to the larger scholarly community.

Outcomes set forth in original project proposal:

> Emory will aggregate metadata from contributing institutions by providing contributing institutions with data conversion expertise. Emory will convert and import existing metadata, in the form of finding aids, catalog records, or other machine-readable forms.

The project has produced many metadata providers in the course of the project, either by directly converting source metadata into OAI-compliant provider systems, or by catalyzing the development of such systems through consultation and advice.

3.1 Data Conversion and OAI Provider Results

As mentioned in the infrastructure section, the OAI-PMH has proved to work well as an interoperable technique that enables pooling metadata.

3.1.1 Provider Systems Produced

A number of OAI data providers have been produced in the course of the project, with a concomitantly large amount of metadata. An appendix to this report provides complete details concerning the providers created, including URL's for the providers, institutions which were the source of the metadata, the associated collection labels, number of records in the providers, and a characterization of the type of records in the providers (granular or collection-level).

As we began to work with other institutions to convert their metadata and make it available for harvesting, we were struck by the variety of metadata forms encountered, and the varying assumptions about metadata that were held by the staff at different institutions. In some cases the information that institutions thought important to disseminate was quite granular, either pertaining to individual artifacts (letters, images, books) or to fairly specific sub-series within a larger collection (boxes or folders of archival materials, typically). In other cases, the primary metadata that institutions wished to share was at a higher level of abstraction, most often descriptions of entire collections, or sometimes collections of collections.

It was also interesting that the resulting providers were sometimes associated with individual collections, sometimes associated with clusters of collections, and sometimes associated with the entire holdings of an institution. There was no obvious trend in how institutions conceptually wanted to relate their providers to their holdings. The results appeared to be mostly based on the most expedient strategy for setting up the provider on the existing digital library architecture, certainly a very pragmatic approach.

Summary of Metadata Provider Systems Produced:

- A total of 23 OAI data providers were created at 16 libraries and archives in the course of the project. These providers were structured as 42 conceptual metadata gateways, as characterized by contributing institutions.
- These providers expose a total of 62,019 records, expressed in unqualified Dublin Core. The source metadata formats included both documented formats such as EAD, MARC, and VRA, as well as ad hoc formats particular to the repository.
- 9 of these 23 providers are operated at Emory University, and 4 of these 9 Emory providers act as surrogate data providers for other institutions.
- The metadata served out by these providers is associated with some 420 different collections.
- Of the 62,019 records exposed, 392 are collection-level records (primarily derived from EAD finding aids), and 61,627 records are more granular (associated with individual items or specific sub-series within archives).
- 16 of the 23 providers were created using the ODL software from Virginia Tech. The remaining providers were ad hoc programming scripts, most often written in PERL.
- 5 of the providers implemented the “sets” feature of the OAI-PMH, in all cases to segment records associated with particular collections.
- The number of metadata records served out by the individual providers ranged in size from over 30K records to a single record.

As mentioned, there were few identifiable patterns in the data conversion streams. Collection level records were as likely to come from static trees of XML as from online catalogs. Granular records originated from both XML and database management systems of various kinds (both networked and standalone desktop systems). Metadata originated in many types of digital library infrastructure (and often traditional library infrastructures, like online catalogs, in which a surprising number of metadata for digital items was being tracked).

There was a consistent collaborative process model for the project data conversion streams involving the smaller MetaArchive institutions (a separate model was used for the larger AmericanSouth institutions). This model was invoked as a way of standardizing the interaction with the smaller MetaArchive institutions. The model was itself a product of the project that we wished to assess, and it worked quite well as a way of structuring and facilitating the creation of metadata from small institutions.

3.1.2 Collaborative Process Model

The central project staff at Emory University developed the following process model to structure metadata identification and collection activities with MetaArchive contributing institutions, beginning with the first test case, Southwestern University. This process organized the work of metadata collection and associated conversion in a conceptual assembly line or stream that was designed to be easily understandable by all stakeholders involved.

The process model was disseminated in various forms during project meetings with partners, and was an important conceptual tool to frame the discussions. Individuals from contributing institutions were often very anxious as the collaborative process began, as they had no frame of reference to understand what a "metadata harvesting" project entailed, or what would be expected of them. The various forms of the process model always described four general phases, with associated work activities broken down into concrete steps:

PHASE I: PREPARATION

- The central project staff would conduct interviews with the partner institution staff to gain an understanding of local metadata aggregations, formats, practices, and applications in use.
- Partner institution staff were presented with guidelines for identifying suitable collections and record.
- The partner institution staff provided the MetaArchive team with access to the relevant metadata. In most cases, the collections were no longer growing, and either a database of holdings or finding aids had been created in the past. This metadata was transmitted to the MetaArchive project staff, most often through network FTP, but sometimes by mailing a diskette or CD-ROM.
- The central project staff performed an analysis of the metadata. If it used a format new to he project team, documentation about the metadata elements would be acquired and studied.
- The central project staff consulted with Emory catalogers to prepare a crosswalk to map the metadata into Dublin Core elements, as well as other potential metadata formats desired by partner institution. The contributing partner staff were often contacted through conference calls to understand the logical structure of the metadata.

PHASE II: CONVERSION

- The crosswalk was reviewed with the partner institution staff. An OAI data provider was established for the metadata at this point, either on an Emory server or at the partner site.
- The partner institution staff sometimes undertook some cleanup of the metadata, in light of inconsistencies revealed.
- Converted metadata was made searchable at a limited-access URL web interface, searchable only by the partner and project staffs.
- The partner staff provided feedback on the converted metadata as well as the search capabilities of the system.

PHASE III: COMPLETION

- The final conversion of the metadata was completed by the central project staff and approved by the partner institution staff. The OAI provider was finalized at this point.
- The newly available metadata was harvested into the union database of all aggregated metadata.

- If the partner institution's collection was still growing or being processed, a schedule for subsequent updates is developed.

PHASE IV: EVALUATION

- Partner institution staff were asked to provide feedback on the process and the results.

The process model worked well on the whole, and we feel it is the most effective structure likely to emerge for these kinds of efforts in which a central project staff work to facilitate smaller institutions in sharing their metadata. However, we observed a number of problems during these collaborations even when using this process model to structure the interaction:

- Explaining the concepts and significance of the OAI-PMH to staff from other institutions was not often easy. This was surprising because the protocol seemed to us to describe a simple and elegant mechanism for sharing information. Throughout the period of 2001 to 2003, however, most individuals we encountered had never heard of it, and often took several discussions for the key ideas to sink in. This problem will hopefully improve as awareness of the protocol spreads. We do think that more efforts to popularize the concept of metadata harvesting are needed to further jumpstart adoption of the protocol.
- It was sometimes difficult to convince staff at other institutions of the potential utility of the OAI-PMH and getting their buy-in. The vast majority of individuals were ultimately either convinced or willing to try out the concept as long as we were doing the work for them. However, two institutions (SMU and Princeton Theological) withdrew from the project because of reservations about the metadata harvesting concept and/or lack of interest.
- In many instances, archival staff were so overcommitted that even when they were interested in the project they genuinely lacked the minimal time required for interacting with our staff. In these cases, we allotted more time to the process and ultimately completed our work with them at a slower pace.
- Finally, a large number of problems emerged associated with the variation in the metadata exposed. These issues are studied in more detail in the next section.

A key question of the project was how a central project team might facilitate wider dissemination of metadata about hidden collections through the OAI-PMH. We feel that the process we engaged in was a successful demonstration that a relatively small project team can facilitate significant metadata dissemination among institutions spread over a broad geographic range. The process is not without difficulties, but worked remarkably well in retrospect.

3.1.3 Analysis of Resulting Metadata Aggregation

The resulting aggregation of metadata was better than we expected in some ways, and worse than we anticipated in others. The variety of obscure collections exposed through the efforts of MetaArchive and AmericanSouth were intriguing. The sheer flexibility of the protocol in enabling information about an unprecedented number of collections and items arising in very localized contexts, but holding great potential for research was astonishing. Browsing and searching the resulting union database is a peculiar kind of activity, with precisely the sense of delving through a vast cabinet of infinitely variable little treasures, obscure troves, and unexpected curiosities. The serendipitous discoveries and connections that one can make between the pied collections assembled here in one place are almost unlimited. The union database also feels different from web search engines. Because the majority of the records aggregated in our database arise from scholarly repositories, the things one encounters in it are authentic objects for research, not advertising gimmicks. But the variability of the items harvested is also a problem. The database

still lacks the sense of order that arises from browsing a library catalog, with the serried ranks of subject headings and uniform headings one finds in the typical OPAC.

We have come to believe that this variability is endemic to first generation metadata harvesting systems. In previous project reports we termed the underlying problem "metadata format collisions". We now have slightly recast the problem as "metadata context collisions."

A sidebar: On the popular CBS television show *C.S.I.: Crime Scene Investigation,* William Petersen stars as Gil Grissom, a senior forensics officer in Las Vegas who leads a team of crime scene investigators. One of Grissom's favorite quotes that he often invokes is: "Evidence without a context is ambiguous at best." After two years of working with OAI metadata harvesting, we similarly submit that *metadata without a context is ambiguous at best.*

Metadata context collisions occur when metadata is aggregated from sources that are not otherwise brought together typically. Metadata is always produced in some relatively well understood and delimited context of practice. The OAI-PMH allows metadata arising from all manner of contexts to be merged without regard for the differences in the original context. This has typically not happened in the past because metadata is expensive to produce, and it is usually held in silos of strictly constrained contexts, with delimiting structures and framing expectations by both end-users and metadata producers. The OAI-PMH is in some ways like a magic key that unlocks the silos, allowing the metadata inside the silos to get out and commingle. This is good in that it allows combinations of information that were never possible before, but bad because it is cognitively disorienting to researchers whose expectations have been framed by siloed systems.

We have observed many sorts of context collisions:

- Collection-level records are usefully produced at a high level of abstraction to provide a summary description of a body of material, or there is not enough staff time to provide detailed information about the collection. When they are encountered in an undifferentiated way with item-level records that are much more granular in scope, researchers are inevitably confused.
- Especially when collapsed from a complex finding aid into a single Dublin Core record, a collection level record may be absurdly more complex than a sparse item level record. The extreme examples of Dublin Core records are those derived from detailed EAD finding aids and those derived from one line descriptions of photographs.
- Subject headings inevitably vary between archives. Any focused collection of primary source materials (say, an archive of an individual's personal papers) will almost certainly require a specialized vocabulary to describe its contents. Focused collections are by definition far more specific than general library collections of books. Such specialized vocabularies are necessary and effective in the context of a focused archive, but hopelessly and uselessly parochial when the records are extricated from a limited context and pooled with records describing all sorts of other archives. Browsing related records by following subject heading links becomes impossible.
- Quite apart from subject headings, cataloging practices of many kinds naturally vary between archival institutions. And why not? The collections are inherently unique; shared cataloging practice is not an issue since the artifacts and therefore their metadata are uniquely bound to the single institution. When records from various archives are pooled, they are inevitably dissonant and disharmonious.
- Even if an institution thinks ahead to the task of converting its specialized metadata records into the unqualified Dublin Core format, the opportunities for variation are enormous. Institutional metadata harvested by the MetaScholar, OAIster, and Illinois metadata harvesting projects have been observed to demonstrate enormous variation in format and usage of virtually all Dublin Core elements [Halbert, 2003].

The problems introduced by metadata context collisions inevitably arise in situations where metadata is aggregated from many different contexts. The problem can conceivably be addressed in at least three ways:

- The differing contexts might be brought into alignment. Thinking back to a time (not so long ago in historical terms) when OCLC and MELVYL were new, the same issue of metadata context collisions was in evidence as cataloging records from different institutions were pooled into union catalogs. Peer-pressures ultimately aligned community practice. This may not be possible in metadata aggregation networks encompassing institutions that have no mandate or directly coupled connections to force alignment of their metadata practice. This is evidenced not only in the Mellon metadata harvesting projects but also in other similarly broad endeavors such as the Semantic Web project of the W3C. In these cases, the potentially commingled sources are likely too numerous and decoupled to be brought into alignment.
- The different contexts might be remediated post-harvest. Experimenting with mechanisms for automatically imposing and/or surfacing latent metadata structures through semantic clustering techniques is the goal of the MetaCombine project, and many others similar efforts such as Data Mining that attempt to organize heterogeneous information after the fact of aggregation.
- Researchers might learn to cope with the lack of uniform context. Especially for focused groups of researchers that had a shared need for aggregating some heterogeneous collection of metadata, a tolerance and understanding of mixed metadata contexts might be acquired. This would depend on the degree of shared context the group could develop in understanding the *lack* of coherent context in their metadata aggregation.

Our findings concerning metadata context collisions were unexpected and troubling, but also stimulating and full of opportunities. We hope to make progress in understanding each of the three possibilities above through future projects.

Considering the first of the three possibilities listed above reoriented our perspective on a larger series of questions about how metadata is not only harvested but produced in the first place. Thinking about the cycle of metadata production and consumption led us to the idea of "metadata gardening."

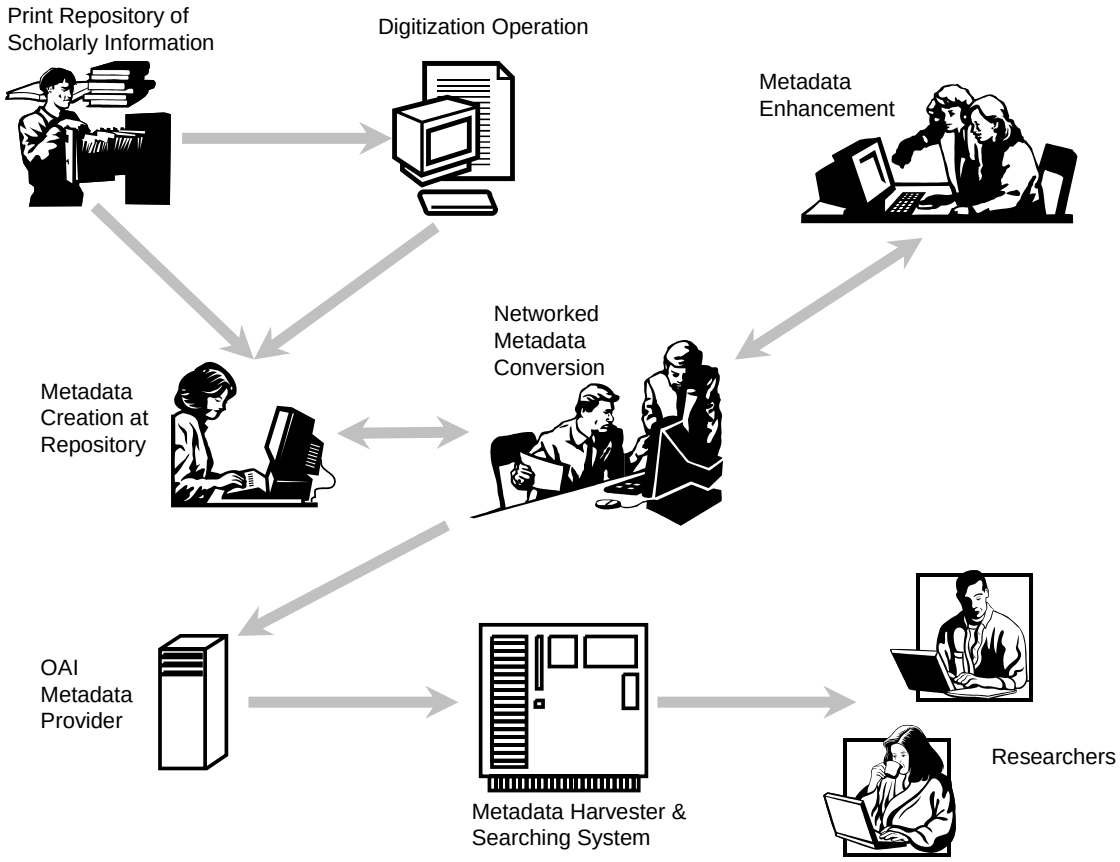
3.2 Metadata Gardening Model

Our experience has led us to conclude that what we were really studying in this project was not so much metadata harvesting as **metadata gardening**, or the entire cycle of metadata creation, dissemination, and consumption that occurs whenever an attempt is made to share metadata across institutional boundaries.

One cause of metadata context collisions is precisely the lack of attention to the entire process of metadata gardening, and the focus on just one part of the cycle: the harvest. If the entire sequence is taken into account from the beginning, the issue of contextual variation becomes clear immediately. It also becomes clear how many extended connections there are in the entire information ecology under consideration.

We examined the major links in the chain of metadata production that were feeding our projects and realized that there were key interactions and loops apparent in the chain. We feel that gardening is a better metaphor for operations such as ours, because we tried to address the activities in other parts of the cycle, and not just the harvesting portion. The following figure attempts to portray these linkages.

Figure 3. Metadata Gardening Cycle



Considering the entire cycle of metadata gardening helped us model the different aspects of the results we were trying to accomplish. Clearly much more study of the motivations and means of mobilizing efforts in this cycle is needed before it becomes commonplace to think about metadata gardening operations. But we hope to have contributed to this understanding through our projects.

4. Scholarly Communication Study

The reason for exploring the new approach to metadata aggregation was the hope that it would produce an information resource of benefit to scholars served by libraries, archives and other repositories of cultural information. We therefore sought to assess the potential utility of the MetaArchive system for scholars through direct feedback from scholars.

Outcomes set forth in original project proposal:

> Emory will evaluate the benefits to the scholarly community at large of this metadata-harvesting infrastructure by several means. Online surveys will be maintained on the web interface. Additionally, focus groups of a) scholars and b) metadata providers will be interviewed for feedback on the service, and suggestions for future directions.

We originally intended to deploy online surveys, but got strongly negative comments from users about such an approach for garnering feedback. Especially for a beginning service that had not yet built up a significant clientele, online surveys seemed to be premature and off-putting to potential users. We instead used small group sessions to get feedback directly from our potential users.

4.1 Focus Group Results

Three focus groups were conducted by MetaArchive staff in April 2002 in order to get direct feedback from potential users. We brought in individuals from various institutions, representative of several categories of potential users, including: tenured faculty, graduate students, librarians, archivists, library administrators, and visiting scholars. In each session, attendees were given background information on the project, a demonstration of a beginning prototype system, and engaged in structured feedback discussions. Five to seven individuals participated in each group.

All of the participants felt the service as demonstrated could be helpful for scholarly research in focused subject domains. Many participants asked questions about how the service would ensure that the metadata would be accurate and trustworthy. It was interesting to note that participants were much more technically astute in the use of online resources than had been assumed, and asked for many additional systems features that we had not anticipated. Some examples of features requested included the ability to browse subject headings and time periods.

Different categories of researchers had quite different expectations for the portal service. Graduate students expressed interest in resources including: contextualizing guides, research methodology information, and pedagogical aids to using the primary sources the service indexed. Some senior researchers were much more resistant to including such features, being skeptical of all preconceived interpretations of information.

One of the scholars attending the focus group sessions was the head of an American Studies department in a German university. Fortuitously, his particular subject interest was the American South, and he had a great number of observations to share about the value of a service which would enable him to canvass the holdings of a number of archives in advance of planning trips across the Atlantic.

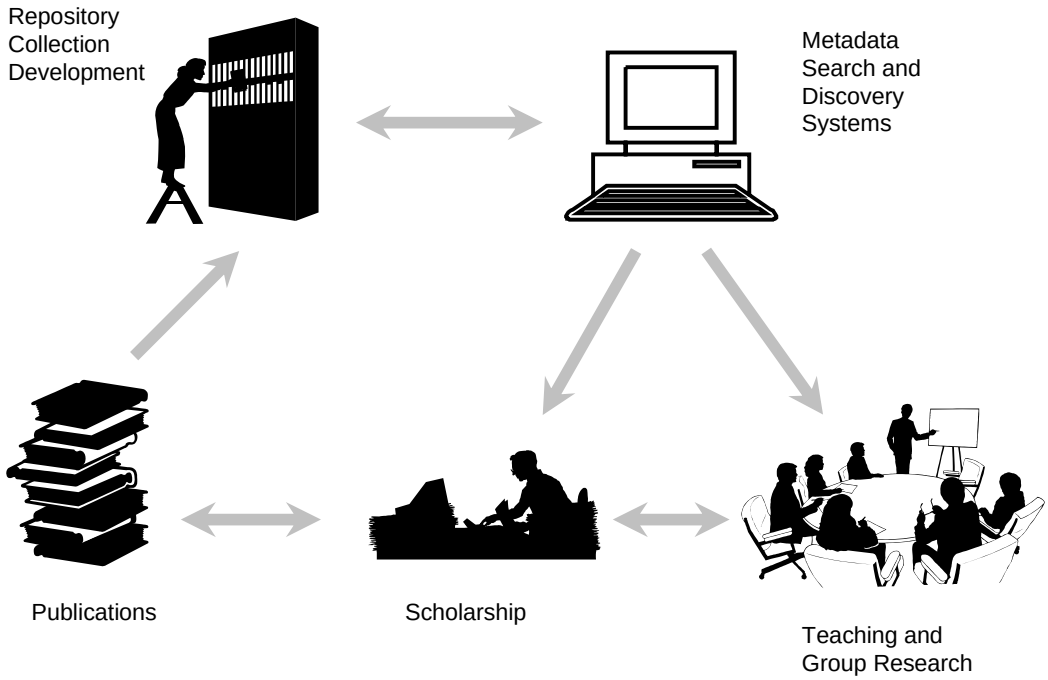
We also heard from a number of scholars that they were interested in simultaneously being able to search many other sorts of information, for example, web pages and peer-reviewed full-text content. This dovetailed with other discussions going on in the MetaScholar Initiative.

In addition to the early MetaArchive focus groups, the AmericanSouth Scholarly Design Team (SDT) functioned as an in depth and ongoing focus group providing feedback and guidance as we experimented with various scholarly communication portal concepts. The extensive thought devoted to project concepts by the SDT is detailed in the AmericanSouth project final report.

4.2 Findings on Metadata and the Scholarly Communication Cycle

There were several unexpected points brought out through the feedback from potential user groups. One recurrent observation was that if we wanted to develop a service that had utility for scholarly communication, we again needed to step back and analyze the many ways that metadata was already used in scholarly communication, and the relationship of metadata to other key activities that affected such communication. These high-level connections are outlined in the following figure.

Figure 4. Metadata and Scholarly Communication



This was an interesting exercise, and reinforced the conclusions from the Subject Portal Working Group about how what we were trying to create was in fact a digital library focused on a subject domain, with systematic connections to scholars in that subject domain, and linkages back to collection development processes.

This reinforced our thinking that we needed to closely connect our efforts with potential users. At he same time, we needed to act as a testbed for new ideas that were not occurring elsewhere. We have tried to incorporate this concept of a close working relationship with scholars into the design of the MetaCombine project, in which an advisory board of scholars is being convened to not only engage in systematic feedback about the clustering experiments of that project, but also serve directly as an editorial board for the new Southern Spaces peer-reviewed forum we are developing based on lessons from AmericanSouth and MetaArchive.

4.3 Analysis of Potential Users

As a result of the focus groups and other discussions conducted in the course of the project, we refined our thinking about the potential users of a scholar's portal built around metadata harvesting capabilities. There are several audiences that might particularly benefit from the kinds of services we have developed.

Figure 5. Some Particular Audiences for Aggregated Metadata from Repositories of Scholarly Information

Librarians, Archivists, & Curators
Interested in:
-Disseminating metadata
-Collection development coordination
-Metadata correction & enhancement activities

Archival Researchers
Interested in:
-Particular subject areas
-Comparing holdings
-Finding particular items

Dissertation Scholars
Interested in:
-Finding particular primary sources for research projects
-Identifying patterns of archival materials held

University Professors
Interested in:
-Assigning primary source research tasks to students
-Identifying materials for curriculum planning

K-12 Community
Interested in:
-Providing K-12 students with an understanding of archives and primary source materials
-Finding materials for cultural heritage appreciation learning

Distant Scholars
Interested in:
-Researching archival holdings in preparation for research trips
-Searching holdings of specialized repositories

In coming projects we will attempt to engage these particular audiences through the creation of targeted services and marketing efforts, as each group has different needs, expectations, and interests. Different discovery and filtering systems will offer utility to each group. In the course of this project we have only been able to catalog these groups, but hope to differentially address them subsequently.

We will investigate new possibilities for organizing information post-harvest for particular audiences in the MetaCombine project, as well as trying out the separate activity of creating new forums for peer-reviewed online scholarship in Southern Spaces. We hope to triangulate on the process of cultivating online communities of scholars through these approaches.

5. Sustainability Study

Because metadata harvesting services are a very new kind of undertaking, there is great need for study of how to sustain such operations. The MetaArchive project sought to understand the options for sustaining such a service, both as a potential mechanism for continuing the particular activities initiated at Emory University, but also to understand how other such services could be developed and supported at other institutions.

Outcomes set forth in original project proposal:

> Emory will produce an assessment of the initial and ongoing costs of this service, as well as potential means of sustaining the service. This assessment will analyze 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. A question for careful examination is the amount of work and costs associated with maintaining a network of institutions contributing metadata. It is our general hope that the evidence will demonstrate that the service is sustainable and of clear benefit to the scholarly community.

5.1 Assessment of Costs

We start by identifying different possible levels of operation and categories of cost. There are many potential approaches to sustaining a metadata harvesting service, depending on the goals of such a service. We have tried to think broadly about this question, abstracting our specific experiences to understand the more general situation.

Clearly, there are many possible goals and potential levels of operation for metadata harvesting services aimed at surfacing hidden scholarly resources. Such services may not be limited to metadata harvesting, as there are many complementary types of knowledge production activities. For instance, our findings to date in the MetaScholar Initiative indicate that metadata harvesting services may also be associated with metadata dissemination, or content production services. Also, while the OAI-PMH is in our view the premier mechanism for disseminating and harvesting metadata, other transmission protocols are of course possible (and have been used in the past).

Analysis of the costs of setting up digital library projects suggests three levels of operation, occurring at logical break points as the scale and duration of the operations increase:

- **Minimal:** A minimal metadata harvesting operation is one with extremely limited goals, operated by 1-2 individuals, with little or no institutional support or organizational structure. Such an operation might be undertaken informally on behalf of a small group of researchers with close ties, interested in some quite specialized topic (for example, members of a research project working on separate streams of investigation, who needed a means of pooling their data collection results and research findings). Support for such an operation would likely entail a part-time commitment of time, and a minimum system infrastructure (perhaps one networked personal computer configured as a server). The planning horizon for maintaining a minimal operation might typically be expressed in terms of months, with focus on only one project of 12-36 months. Examples of a minimal operation might include a group of nationally collaborating scholars, or a collaboration among a small number of libraries.
- **Median:** A medium-scale metadata harvesting is one with more extended goals, operated by 3-10 individuals, with modest institutional support (for example, office space provided) and operating as part of an existing organizational structure. This level of operation might be

undertaken with semi-formally, on behalf of a moderately large group of researchers with more tenuous connections through professional groups or institutional sponsorship, interested in a cluster of conceptually related topics (for example, the research in a newly emerged academic discipline, who are interested in sharing and identifying findings as new publications arise). Support needed for this level of operation would require a mix of full and part-time commitments on the part of various individuals depending on their roles, and a system infrastructure of several servers, some of which are upgraded and replaced over time. The planning horizon of a median operation might be projected through a series of related projects over 2-6 years. Examples of median-scale services might include the MetaScholar Initiative at Emory University or operations such as the Perseus Digital Library or the UVA E-Text Center (although the two latter operations are not aimed at metadata harvesting they give an additional sense of scale and the typical connections with a parent institution).

- **Advanced:** An advanced metadata harvesting is conceivably a very large organization of 10+ individuals that may constitute a dedicated institution of their own. This scale of operation would have formal agreements that it entered into with other institutions and associations, on behalf of very large groups of researchers with no direct connections who were interested in many wide-ranging areas of research (for example, established academic disciplines). The support for this scale of organization would require numerous full-time staff. The systems infrastructure would be extensive, with dedicated servers addressing various categories of functions in the metadata gardening cycle, including: harvesting, dissemination, record enhancement, publication, or digital processing. The infrastructure of such an organization will require significant ongoing investment, and planning horizons might extend out to a decade or more, with ongoing programmatic planning by senior managers. Some possible examples of advanced operations might include OCLC and RLG.

There are two general types of costs involved in any service: one-time and ongoing expenses. For this analysis, we will assume that open source software is used, entailing no direct expense. The following analysis breaks down these costs by the different levels of operation described above using several generic cost estimates:

$56,000	Typical professional salary (this is the typical salary for an average faculty member or a senior librarian, using figures from the current *Occupational Outlook Handbook*.)
$70,000	Typical annual cost of a professional, counting salary plus 25% benefits rate.
$35/hr	Typical hourly cost of professional labor, assuming 2000 working hours in the year, or 40 hrs/week x 50 working weeks/year x 1 year. In academic settings this amount might be lower because of extended leave time.
$5,000	Typical cost of a small-scale computer usable as a server

5.1.1 One-time Costs

Estimated typical one-time costs for different operational levels, based on our experience in this project, and extrapolating for smaller and larger operations, include:

Minimal Operations (and above):

- Initial planning, research, and familiarization with metadata formats and harvesting systems infrastructures that will be used (480-1,440 hrs labor).
- Hardware expenses (est. $5,000-$10,000).
- Development of the metadata gardening infrastructure, including provider and harvester systems (480-1,440 hrs labor).

- Development of other related systems for communication and other processing (480-1,440 hrs labor).
- Testing, revisions, documentation, and other follow-up to initial deployment (480-1,440 hrs labor).

Median Operations (and above):

- Service group formation activities, including recruiting, orientation, and training (80-240 hrs labor).
- Additional hardware expenses (est. $10,000-$20,000).
- Extended infrastructure development (160-480 hrs labor).
- Initial planning with service clientele to meet and develop the service (480-1,440 hrs labor + $5,000 meeting/travel expenses).

Advanced Operations:

- Service organization formation activities, including programmatic planning, recruiting, orientation, and training (4,320-12,960 hrs labor).
- Facilities procurement (est. $100,000-$200,000).
- Startup Publicity and promotions (est. $50,000-$100,000).

5.1.2 Ongoing Costs

Estimated typical ongoing costs for different operational levels, based on our experience in this project, and extrapolating for smaller and larger operations, include:

Minimal Operations:

- Minimal operations by definition are short term and focused on individual projects. If they continue into subsequent projects they will most likely evolve into median level operations.

Median Operations:

- Annual salaries of 3-10 individuals ($210,000-$700,000 per year).
- Annual hardware expenses (est. $5,000-$20,000).
- Annual meetings with service clientele to continue development of the service ($5,000-$20,000 meeting/travel/other expenses).

Advanced Operations:

- Annual salaries of 10-40 individuals ($700,000-$2,800,000 per year).
- Other annual operating expenses ($350,000-$1,400,000 per year).

Figure 6. Estimates of One-Time and Ongoing Costs

Service Operation Costs		Low	High
One-Time Expenses			
Minimal		**$ 72,200.00**	**$ 211,600.00**
Initial planning,research, and familiarization		$ 16,800.00	$ 50,400.00
Development of the metadata gardening infrastructure		$ 16,800.00	$ 50,400.00
Hardware costs		$ 5,000.00	$ 10,000.00
Development of other related systems		$ 16,800.00	$ 50,400.00
Testing, documentation, revisions, and other follow-up		$ 16,800.00	$ 50,400.00
Median (Minimal + Additional)		**$ 123,600.00**	**$ 345,800.00**
Service group formation activities		$ 2,800.00	$ 8,400.00
Additional hardware expenses		$ 10,000.00	$ 20,000.00
Extended infrastructure development		$ 16,800.00	$ 50,400.00
Initial planning with service clientele		$ 21,800.00	$ 55,400.00
Advanced (Median + Additional)		**$ 424,800.00**	**$ 1,099,400.00**
Service organization formation activities		$ 151,200.00	$ 453,600.00
Facilities procurement		$ 100,000.00	$ 200,000.00
Publicity and promotions		$ 50,000.00	$ 100,000.00
Ongoing Expenses			
Median		**$ 220,000.00**	**$ 740,000.00**
Annual salaries of 3-10 individuals		$ 210,000.00	$ 700,000.00
Annual hardware expenses		$ 5,000.00	$ 20,000.00
Annual meetings with service clientele		$ 5,000.00	$ 20,000.00
Advanced		**$ 1,050,000.00**	**$ 4,200,000.00**
Annual salaries of 10-40 individuals		$ 700,000.00	$ 2,800,000.00
Other annual operating expenses		$ 350,000.00	$ 1,400,000.00

The figures above are reasonable estimates for the general scale of one-time and ongoing expenses of the sort of services examined in the MetaArchive project. In the next section we will consider models for sustaining such services financially.

5.2 Options for Sustainability

There are a large number of conceivable business models for sustaining a service providing metadata harvesting and other complementary functions for scholarly communities. This topic is closely related to the more general question of sustaining emerging digital library services. This issue has received much attention in recent years as such services proliferate in testbed operations that subsequently seek to become mainstreamed.

A prominent service that is currently grappling with this question is the National Science Digital Library (NSDL), a relatively large digital library service created by more than $100M in federal research grants issued by the National Science Foundation (NSF) in the last few years. In the context of the three-level taxonomy developed in the previous section, the NSDL is a good example of a very advanced operation, aiming at great breadth as a multi-disciplinary metadata harvesting and content publication portal. The start-up costs of this project were much larger than the estimates given in the previous section. The NSDL is now in the process of identifying a model for sustaining its operations, and has considered a variety of business models in a recent white paper [McArthur, 2003], that are worth reviewing here.

NSDL Business Options Considered:

- Open-access aggregator/distributor of free resources that require no direct payments from end-user communities.
- Partner with commercial or professional society publishers.
- Value-added service provider for specific sectors such as K-12.
- Institutional subscriptions.
- Research facility, sponsored by additional funding from NSF.
- Self-sustaining community or set of communities, similar to SlashDot or Wikipedia.

Each of these models might also act in combination with one or more of the others to good effect. In considering these options, the NSDL has also identified a framework of questions for analyzing these options. Questions being asked include things like: *Why is it important? Who will want it and pay for it? What is the comparative advantage of NSDL?*

Thinking back to the three levels of operation described previously, what becomes clear is that each scale of operation has different sensible funding models. In a minimal operation, basic operational funding comes primarily from some parent organization such as an academic department. The metadata harvesting operation is probably seen as very speculative and ad hoc funding needed to undertake these speculative activities may be received through either a small research grant or discretionary allocation of time by interested researchers. The metadata aggregation function is only an ad hoc means to larger programmatic goals, and is not a critical service that need be sustained over time.

When an operation grows to a median level, it begins to assume some programmatic aspects in its own right. Undertaking this scale of operation requires explicit ongoing operational funding to sustain the activity. Although sources of funding may still primarily come from a parent institution and research grants, additional options such as commercial services (consortia subscriptions, fee-based consultations, professional society dues, or advertising) or governmental agency roles become options as well.

When an operation grows to an advanced level, it almost certainly becomes a separate enterprise because it is too large to avoid overwhelming most parent institutions such as an academic department or research library. This has several implications. First, it is unlikely to be able to generate enough revenue to sustain itself through anything but either commercial services or becoming a governmental agency.

This progression seems to suggest the likely models for sustainability of operations of these three scales, and provides a practical set of options for the MetaScholar Initiative projects as a case study. Categories of funding that our projects have received to date include research grants from foundations such as Mellon and federal grant agencies like IMLS, as well as basic institutional support from Emory University. In future, we may consider additional funding sources, for example consultation services for institutions seeking to set up metadata provider systems. We have considered this option, and a conceptual example of a consulting services schedule is included in the appendix. We do not intend to offer subscription services, as this would be in conflict with our commitment to open access to scholarship.

The most important issue to consider is whether a service like ours offers value to scholars. We think that the continued involvement of the scholars who have participated in the project demonstrates this to some extent, but this is an issue to gauge in coming months as we roll out additional services such as the Southern Spaces online forum and other resources.

6. Conclusions

The following are summary statements of the major project conclusions:

- Effective, scalable infrastructures of both OAI data providers and metadata harvesting operations can be developed using only open source software tools. Such systems can be deployed with low barriers to entry by small to large operations.
- Metadata harvested from distributed OAI data providers and made keyword-searchable through web interfaces is a valuable scholarly resource with several potential audiences. However, scholars are also interested in browsing such information in ways not currently possible because of the problem termed metadata context collisions. Scholars are also interested in simultaneously being able to search a number of other sorts of information, for example, web pages and peer-reviewed full-text content.
- There may be ways of remediating the problems associated with metadata context collisions, either through pre-harvest alignment of metadata production contexts, or post-harvest processing through semantic clustering approaches. We intend to investigate both of these options in the future.
- There seem to be several logical approaches to sustaining projects such as this one over time. Various combinations of funding strategies present themselves, each with a different dynamic depending on the scale of the operation

7. References

[Campbell, 2000] Campbell, Jerry. "The Case for Creating a Scholars Portal to the Web: A White Paper." *ARL Bimonthly Reports*, Issue 211, August 2000. URL: http://arl.cni.org/newsltr/211/portal.html

[Halbert, 2003] Halbert, Martin, et al. "Findings from the Mellon Metadata Harvesting Initiative." Research and Advanced Technology for Digital Libraries, 7th European Conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003, Traugott Koch, Ingeborg Sølvberg (Eds.), Proceedings. Lecture Notes in Computer Science 2769 Springer 2003, ISBN 3-540-40726-X. pp. 58-69.

[Jones, 2003] Jones, Barbara, et al. *Hidden Collections, Scholarly Barriers: Creating Access to Unprocessed Special Collections Materials in North America's Research Libraries.* White Paper for the Association of Research Libraries Task Force on Special Collections. Given as part of a Workshop held September 8-9, 2003 at the Library of Congress, Washington, D.C. URL: http://www.arl.org/collect/spcoll/ehc/HiddenCollsWhitePaperJun6.pdf

[Maly, 2003] Maly, Kurt. "ARC – An Open Source Metadata Harvesting System." *Workshop on Applications of Metadata Harvesting in Scholarly Portals*, Emory University, Atlanta, Georgia. October 24, 2003. pp. 36-40.

[McArthur, 2003] McArthur, David, et al. "A Summary of Business Options for NSDL." National Science Digital Library, internally circulated document, 10/10/2003.

8. Appendices

Two appendices are provided as attachments.

Details on OAI Data Providers Created

This appendix provides details concerning the OAI data providers created, including URL's for the providers, institutions which were the source of the metadata, the associated collection labels, number of records in the providers, and a characterization of the type of records in the providers (granular or collection-level).

Example of Metadata Consulting Services Schedule

As an exercise, a schedule of metadata consulting services was developed for distribution to other institutions interested in contracting with our group to develop OAI data providers. It provides a potential example of other means of generating funds for sustainability.

Appendix: Details on OAI Data Providers Created

	Institution	Collections	Record Count	Record Type	Provider URL	Set
1	Atlanta History Center	Atlanta History Center	2900	Granular	http://callimachus.library.emory.edu/cgi-bin/oai2_providers/XMLFile/atlantahistorycenter/oai.pl	
2	Auburn Univeristy	AAA Finding Aid Series	5	Collection	http://diglib.auburn.edu/cgi-bin/OAI-XMLFile/XMLFile/auburn/oai.pl	AAA
3	Auburn Univeristy	ABH Finding Aid Series	1	Collection	http://diglib.auburn.edu/cgi-bin/OAI-XMLFile/XMLFile/auburn/oai.pl	ABH
4	Auburn Univeristy	ABJ Finding Aid Series	3	Collection	http://diglib.auburn.edu/cgi-bin/OAI-XMLFile/XMLFile/auburn/oai.pl	ABJ
5	Auburn Univeristy	ACM Finding Aid Series	2	Collection	http://diglib.auburn.edu/cgi-bin/OAI-XMLFile/XMLFile/auburn/oai.pl	ACM
6	Auburn Univeristy	ALM Finding Aid Series	1	Collection	http://diglib.auburn.edu/cgi-bin/OAI-XMLFile/XMLFile/auburn/oai.pl	ALM
7	Auburn Univeristy	AVY Finding Aid Series	2	Collection	http://diglib.auburn.edu/cgi-bin/OAI-XMLFile/XMLFile/auburn/oai.pl	AVY
8	Auburn Univeristy	Transforming America	213	Collection	http://diglib.auburn.edu/cgi-bin/OAI-XMLFile/XMLFile/aufindaid/oai.pl	
9	Davidson University	Ney,Sayyid	5	Collection	http://callimachus.library.emory.edu/cgi-bin/oai2_providers/XMLFile/davidson/oai.pl	
10	Emory University	American Routes	144	Granular	http://callimachus.library.emory.edu/cgi-bin/oai2_providers/XMLFile/amroutes/oai.pl	
11	Emory University	Ralph McGill	20	Granular	http://sage.library.emory.edu/cgi-bin/OAI-XMLFile/XMLFile/SAGE/oai.pl	Ralph_McGill
12	Emory University	Southern Changes	52	Granular	http://chaucer.library.emory.edu/cgi-bin/OAI-XMLFile/XMLFile/schanges/oai.pl	
13	Emory University	Sam Nunn	697	Granular	http://sage.library.emory.edu/cgi-bin/OAI-XMLFile/XMLFile/SAGE/oai.pl	Sam_Nunn
14	Kentucky Virtual Library	Kentuckiana Digital Library	12	Collection	http://kdl.kyvl.org/cgi-bin-oai/XMLFile/kydl/oai.pl	
15	LSU	American South	1739	Granular	http://www.lib.lsu.edu/cgi-bin/OAI-XMLFile/XMLFile/AmerSouth/oai.pl	
16	Pitts Theology	Manuscripts of Religious Writings	262	Granular	http://callimachus.library.emory.edu/cgi-bin/oai2_providers/XMLFile/pitts/oai.pl	
17	Southwestern	John Tower	18186	Granular	http://callimachus.library.emory.edu/cgi-bin/oai_providers/tower/scripts/oaicgi.pl	
18	UNC	Library of Southern Literature	57	Granular	http://www.lib.unc.edu/cgi-bin/oai/das/das/das/oai.pl	2
19	UNC	NC and the Great War	146	Granular	http://www.lib.unc.edu/cgi-bin/oai/das/das/das/oai.pl	1
20	UNC	NC Experience	274	Granular	http://www.lib.unc.edu/cgi-bin/oai/das/das/das/oai.pl	6
21	UNC	Slave Narrative	230	Granular	http://www.lib.unc.edu/cgi-bin/oai/das/das/das/oai.pl	3
22	UNC	Southern Homefront	403	Granular	http://www.lib.unc.edu/cgi-bin/oai/das/das/das/oai.pl	4
23	UNC	Church in Southern Black Community	1957	Granular	http://www.lib.unc.edu/cgi-bin/oai/das/das/das/oai.pl	5
24	United Methodist Archives	Methodist Church Records	97	Collection	http://www.gcah.org/cgi-bin/OAI-XMLFile/XMLFile/OAIgcah/oai.pl	
25	University of Florida	Florida Environment Online	474	Granular	http://brokert8.fcla.edu/cgi/b/broker/broker	amsouth:feol
26	University of Florida	Florida Heritage	1394	Granular	http://brokert8.fcla.edu/cgi/b/broker/broker	amsouth:fhp
27	University of Florida	Florida Historical Legal Documents	6	Granular	http://brokert8.fcla.edu/cgi/b/broker/broker	amsouth:law
28	University of Florida	Florida Maps	107	Granular	http://brokert8.fcla.edu/cgi/b/broker/broker	amsouth:mapfl
29	University of Georgia	Digital Library Collection	905	Granular	http://dlg.galileo.usg.edu/cgi-bin/OAI-XMLFile/XMLFile/senad/oai.pl	
30	University of Georgia	Russell Collection	45	Collection	http://tennessee.cc.vt.edu/~rkelapur/cgi-bin/uga/IB.pl	
31	University of Tennessee	Cherokee	166	Granular	http://helios.dii.utk.edu/cgi-bin/oai.cgi	che
32	University of Tennessee	Civil War	1	Granular	http://helios.dii.utk.edu/cgi-bin/oai.cgi	civ
33	University of Tennessee	Emancipator	7	Granular	http://helios.dii.utk.edu/cgi-bin/oai.cgi	emn
34	University of Tennessee	Great Smokey Mountains	49	Granular	http://helios.dii.utk.edu/cgi-bin/oai.cgi	gsm
35	University of Tennessee	Roth Photography Collection	88	Granular	http://helios.dii.utk.edu/cgi-bin/oai.cgi	rth
36	University of Tennessee	Tennessee Document History	37	Granular	http://diglib.lib.utk.edu/oai/oai20.php	tdh
37	UVA CDH	Virtual Jamestown	67	Granular	http://www.vcdh.virginia.edu/oai/OAI-XMLFile/XMLFile/jamestown/oai.pl	
38	UVA CDH	Valley of the Shadow	1029	Granular	http://www.vcdh.virginia.edu/oai/OAI-XMLFile/XMLFile/valley-letters/oai.pl	
39	Virginia Tech	Imagebase	30230	Granular	http://uther.dlib.vt.edu:8080/oaicat/servlet/OAIHandler	
40	Washington & Lee	Robert E. Lee	2	Collection	http://callimachus.library.emory.edu/cgi-bin/oai2_providers/XMLFile/washingtonlee/oai.pl	
41	Washington & Lee	Dupont	2	Collection	http://callimachus.library.emory.edu/cgi-bin/oai2_providers/XMLFile/washingtonlee/oai.pl	
42	Washington & Lee	Whitehearst	2	Collection	http://callimachus.library.emory.edu/cgi-bin/oai2_providers/XMLFile/washingtonlee/oai.pl	

Metadata Consulting Services of the MetaScholar Initiative

A Brief Guide for Libraries and Archives

This handout describes various metadata consulting services offered by the MetaScholar Initiative, a digital library services program based at Emory University in Atlanta, Georgia.

What is Metadata?

In this context, metadata is any type of information about research collections held by libraries, archives, museums, and other repositories of scholarly information. Metadata may come in the form of cataloging records, finding aids, and other records that describe holdings of research collections.

Benefits of Sharing Metadata through the OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is rapidly becoming the de facto standard technology for broad, public dissemination of metadata through automated means. By creating access to local research collections through the OAI protocol, institutions can quickly and easily disseminate information concerning holdings to many new indexing services that utilize the OAI-PMH such as the AmericanSouth.Org portal, and the OAIster service maintained by the University of Michigan. The OAI-PMH also allows institutions to flexibly transmit and re-use metadata internally in new ways. For example, records for new holdings accessioned in an archival unit can simultaneously be transmitted to the online catalog and a campus portal. The OAI-PMH has been adopted as the core technology of the National Science Digital Library and a growing number of new discovery services. For more information concerning the OAI-PMH, see the OAI website (http://www.openarchives.org).

What is the MetaScholar Initiative?

The MetaScholar Initiative is an ongoing series of projects developing innovative uses of metadata for scholarly communication. Based at Emory University, the MetaScholar Initiative has staff experienced in the practical aspects of adapting OAI-PMH technologies to a wide variety of library system infrastructures.

MetaScholar Staff have successfully created a large number of OAI compliant systems for the purpose of disseminating research collection metadata. Our staff are capable of serving in many consultative roles, including advice on OAI-PMH issues, design of data provider systems compliant with the OAI-PMH standard.

Consulting Services Offered

The MetaScholar Initiative offers several types of consulting services concerned with metadata technologies. The following is a brief summary of these services and typical fees associated with the service:

Basic Metadata Readiness Assessment: $ 400. This assessment will include a conference call between MetaScholar Staff and the contracting institution, to determine the status of their local system capabilities, what steps would be necessary for developing an OAI-compliant metadata provider system for the institution. The deliverable is a readiness assessment report detailing this information, and general recommendations on how to proceed by the MetaScholar staff. If requested, these recommendations will advise the institution on how many subsequent consulting days by MetaScholar staff would be required to establish an OAI metadata provider system.

Metadata Provider System Development Consulting: $ 800 per day. This service can include any form of consultation necessary to get the metadata provider of the institution up and running, including programming services, metadata crosswalks, export and import scripts, etc.

Surrogate Metadata Provider System Hosting: $ 1,600 per year. For this amount, the MetaScholar Initiative will, for a period of one year, maintain a metadata provider system for the institution, register it with the OAI community, and publicize it to discovery services. The institution will be able to monitor the activity on the metadata provider by means of a web page that displays how many metadata records have been transmitted from the provider.

Contacting the MetaScholar Initiative Staff

If you have metadata consulting needs, please initiate a conversation with our staff. You can do so by contacting:

Martin Halbert (Executive Director, MetaScholar Initiative) 404-727-2204

Carrie Finegan (Administrative Assistant, MetaScholar Initiative) 404-712-2024