Proposal to the Andrew W. Mellon Foundation

A Cross-Institutional Scholarly Metadata Service for Harvesting, Searching, and Subject Organization

(excerpts)

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

URL to this document: http://metascholar.org/about/metaarchive/MetaArchive.pdf

Summary

In response to the gracious invitation by the Andrew W. Mellon Foundation, this proposal has been prepared by the Emory University Libraries for a demonstration project and feasibility study of a cross-institutional scholarly portal service providing public search and subject organization functions for metadata aggregated using the new Open Archives Initiative (OAI) harvesting protocol. A timeline and estimated budget is provided.

Date of Submission: Thursday, April 5, 2001.
Proposing Institution: University Libraries, Emory University, Atlanta, Georgia, 30322

Contact Information

Principal Investigator: Martin Halbert,
Director for Library Systems, Phone: 404-727-2204

Additional Contact: Joan Gotwals,
Vice-Provost and Director of Libraries, Phone: 404-727-6861

Project Details

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.

Metadata Conversion Activities. As part of the service, Emory will provide contributing institutions with data conversion expertise. Emory will convert and import existing metadata, in the form of finding aids, catalog records, or other field-delimited machine-readable forms (such as spreadsheets, labeled word processing files, or PC databases). The process for these data conversion activities will include the following steps: 1) detailed interviews (usually by phone, but may include site visits) with personnel from contributing institutions to understand the structure of the metadata that they are contributing, 2) transmission of unprocessed "raw" metadata to Emory, 3) identification of equivalencies between fields in the unprocessed metadata and Dublin Core elements, 4) development of Perl scripts to convert and import the metadata into a staging area of the harvester, 5) confirmation with contributing institutions that the metadata has correctly been represented (accomplished by having the contributing institutions review the metadata in the staging area), and 6) movement of the metadata from the staging area into the harvester's publicly accessible database.

The service aims to not burden contributing institutions or Emory with any additional work processing or re-processing collections, but rather to maximize use of metadata that has already been created for local uses. Archival processing or recon projects are therefore not within the scope of what is proposed, only data conversion. It must therefore be stressed that metadata which is not machine-readable, or comprehensibly parsable will be considered unacceptable for the project's purposes. Emory reserves the right to refuse metadata that cannot be imported through reasonable script-conversion efforts, or which is deemed too poorly structured to be intelligible to scholars. This requirement is made in order to ensure that metadata imported into the harvester will take a useful and effective form, in order to be of value to scholars in search of information.

Notes will be retained during conversion activities concerning what type of factors in the incoming "raw" metadata lead to either difficulties or ease in conversion. It is hoped that these notes may lead to subsequent guidelines for institutions and personnel when creating new repositories of metadata that they may wish to contribute to OAI compliant services. These guidelines, along with lessons learned from the conversion activities, are among the types of information that will be reported in conference and other public presentations on the project.

Why the proposed service is coherent and useful: The service provided will be a coherent and useful tool for scholarly purposes because it will provide scholars with *both* sophisticated search functionality for aggregated metadata *and* organized subject domain presentations. The sophisticated metadata searching capability of the system (keyword, boolean, field/institutional delimitation, etc.) will represent a core utility of the system for scholars. Targeted subject organization areas will be developed as well (see below).

Scholarly Value of Subject Domains. Metadata subject domains of value to scholars will be cultivated from the aggregated complimentary metadata contributed by the various institutions that have agreed to partner with Emory. Two subject domains have been identified so far in the special collections of the contributing institutions (see section on contributors for more details): 1) archived papers of major political figures, and 2) institutional theological records. These two subject areas are deemed to be of scholarly value and worthy of attention in the project for several reasons. These areas represent primary sources that are frequently used in a variety of disciplines in both the humanities and social sciences. This evaluation arises from observation of research activities at Emory and other institutions. The papers of political figures are relevant to research by historians, political scientists, economists, sociologists, or virtually any scholar examining events or policy in the public sphere. Institutional

theological records are relevant to research by theologians, historians, sociologists, or anyone interested in the changing landscape of religious belief. Such materials are frequently used in unanticipated ways; a recent interdisciplinary doctoral project at Emory sought to examine the historical role of women in the Methodist Church through such institutional records. A major problem was simply reviewing what information was to be had at regional repositories without physically traveling to many locations. This sort of research will be greatly facilitated by the proposed metadata harvesting service.

Culling sources is a general difficulty that scholars are frequently observed to encounter. Simply locating *worthwhile* sources of primary materials *relevant* to their research is a considerable component of the research task. By aggregating metadata concerning such archives of primary sources, we hope to provide a means for scholars to be able to simultaneously examine and meaningfully compare multiple primary sources of research material. Where these sources are available in digitized form, an immediate access mechanism will also be provided by means of URLs. Simply aggregating metadata from multiple institutions increases the likelihood of serendipitous encounters with materials that scholars might never have considered if they were searching through materials located at one archive alone.

The two subject domains identified must not be seen as a limiting extent of the scholarly subjects addressed by the project. It is anticipated that additional subject areas will be identified in the course of the project. Criteria for decisions concerning which additional subject areas will be added will be based on whether or not 1) the subject is deemed of research value to scholars, 2) relevant subject expertise exists at Emory to organize a presentation of the material, 3) a critical mass of metadata concerning the subject area is available for harvesting, and 4) no other scholarly portal service already exists for the subject area.

Subject Organization. The Emory staff has gained much relevant expertise in recent years by organizing and developing access to interdisciplinary collections of information residing both at Emory and elsewhere. Staff who have worked on these projects are located in user services, special collections, electronic text centers, statistical data services, and cataloging departments. The subject experts and catalogers listed in the personnel section will act as a working group to cultivate organized and coherent portals for these subject domains during the project.

These subject portals will include several features that will make them valuable to scholars researching a particular topical area. Searching features will enable queries either limited to the subject domain or of broader scope. Hierarchical arrangements of topics, collections, and research strategies will enable scholars to gain an intellectual understanding of the material accessible through the harvesting service and relationships between sources. The subject portals working group will act in concert with the programming team, under the overall direction of the project leader, to develop the subject organization aspects of the system.

Open Source/Free Software Technology: Emory will build on years of experience in creating and maintaining web databases using open source technologies. In previous grant-funded collaborative efforts such as the SAGE Project, Emory has created an infrastructure that can be freely shared with other institutions by basing systems development wholly on open source technologies such as the Linux operating system, the Apache web server, and public domain search engines. Our hope in this project is to offer the benefits of open source technologies to contributing institutions that may not have the resources to maintain such infrastructures. The Emory libraries have demonstrated the utility of new varieties of digital library technologies in recent years (examples include digital archives and full-text center operations), and have successfully transitioned these activities from initial grant-funding status to mainstream budget lines.

Standards: The OAI protocol will be implemented as a mechanism for aggregating metadata automatically from other OAI-compliant repositories. Metadata feeds from contributors partnering with Emory will take place using conversion scripts written in PERL. The Dublin Core metadata fields will form the nucleus for a database structure that will be enhanced with additional fields germane to the resources we will be working with.

Institutional Contributors/Partners

By providing a service for accepting and hosting metadata from contributing institutions of various types, Emory hopes to both benefit scholars through access to information and also build a model for cross-institutional cooperation in surfacing scholarly collections.

Choice of Contributors. Institutional contributors/partners have been identified primarily through two criteria: 1) linkages or collaborative relationships of some form exist between the instituion and Emory, and 2) the institution holds special collections of scholarly value within the two initially identified subject domain areas. Many of the universities identified are members of the Associated Colleges of the South consortium, for which Emory University serves as the administrative host institution. Some of the entitities listed are organizationally part of the Emory University Libraries. Other institutions were approached for a combination of reasons, such as collaborative relationships and known strengths in particular collections.

A variety of institutional sizes and relationships have intentionally been chosen in hopes of learning what kinds of metadata contribution and conversion issues result from various situations. It was felt that choosing a single category of contributor would not lead to the breadth of experience needed to adequately report on the variety of issues likely to be encountered by other subsequent metadata aggregation initiatives. Our contributors include many categories, such as libraries of large state research universities, archives of teaching colleges, special libraries, museums, etc. We feel that this broad spectrum of institutions will provide the needed experience to make generalizations about the process of metadata aggregation.

Timeline

2001 July	Project begins.
2001 Fall	Outcomes: Programming support identified and contracted/ recruited. Initial development work done. Metadata conversion activities for "low-hanging fruit" contributors will take place during this time frame and conclude by the end of the year. Opportunities for presentations concerning the project will be identified at upcoming conferences and professional meetings (possibilities include DLF, CNI, ALA, ACRL, and ASIS).
2002 Spring	Outcomes: Alpha (first working) version of infrastructure in place. Goal of Alpha version is to provide OAI compliant functions, searching, and initial subject domain organization. More complex metadata conversion activities will begin taking place during this time frame and will continue for the rest of the project.
2002 Summer	Outcome: Feedback on Alpha infrastructure collected in structured sessions for Beta design. First year project report to Mellon.
2002 Fall	Outcomes: Beta and subsequent versions of infrastructure released with version control. Goal of Beta and subsequent versions is to refine usability of service. Online user response surveys in place.
2003 Spring	Outcome: Online user response surveys tabulated and reported. Analysis of project costs conducted.
2003 Summer	Outcomes: Feedback on Beta infrastructure collected in structured sessions for final evaluation. Assessment and proposal review to mainstream activity. Final project report to Mellon.