MetaScholar Initiative
Emory University

General Libraries
540 Asbury Circle
Emory University
Atlanta, GA 30322

Phone 404 727 2204
Fax 404 727 0827

MetaCombine Project Proposal

An Experimental Demonstration of Improved Techniques for Organizing Combinations of Metadata and Web Resources for Scholarly Communication Purposes

(excerpts)

July 2003

URL to this document: metascholar.org/about/metacombine/MetaCombine.pdf

Executive Summary

Emory University seeks to conduct practical experimentation with improved techniques for organization and access to scholarly information via the *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) as well as the World Wide Web.

Through the proposed project, Emory will explore combinations of information and services at various levels of abstraction: combined search of OAI and Web resources, combined semantic clusters of information, and combined digital library components acting as a whole. Hence, the project name: MetaCombine.

Key Points

1) The MetaCombine project will assess the effectiveness of several specific *semantic clustering* techniques (see glossary for a description of this approach to information organization) for improving organization and access to bodies of metadata exposed via the OAI-PMH as well as Web resources. The focus will be on two prominent techniques: the *support vector machine* (SVM) class of algorithms and *multidimensional scaling* (MDS) visualization (see glossary for an overview of these methodologies).

2) This project will develop, demonstrate, and assess two approaches for providing combined search capabilities of harvested metadata and Web content: A) by providing OAI access to metadata records automatically generated for Web content via semantic clustering techniques; and B) indexing combined bodies of Web content and OAI metadata.

3) Further, the project will experimentally develop and evaluate a framework for coordinating loosely coupled components of digital library services in an extensible manner, based on a new approach to using the OAI-PMH and Web services as underlying means of system integration.

4) The MetaCombine project will build on the technical expertise and working relationships with scholars accumulated in the AmericanSouth and MetaArchive projects conducted at Emory University.

5) Emory University will use and develop only *open source software* (OSS, see glossary for details concerning this class of software), specifically software that can subsequently be used freely by other research institutions.

6) This project can potentially have broad impacts on many other initiatives, by advancing the current understanding of several areas of digital library technology and scholarly communication.

Problems Addressed

During the course of the Mellon Metadata Harvesting Initiative projects, and generally during the first few years of experimentation with the OAI-PMH, several problems have come into focus. Through the proposed project, Emory seeks to address these problems, summarized below.

- **Problems in Categorizing and Browsing Harvested Metadata:** To date, virtually no OAI-PMH harvesting project has developed effective means of browsing metadata aggregations by subject, author, or other systematic categorization scheme. These problems stem from the fact that metadata aggregations harvested from multiple institutions suffer from a lack of controlled vocabulary and authority control in the underlying *Dublin Core* (DC, see glossary) metadata distributed via the OAI-PMH. Without this consistency, there is no effective way to browse metadata across institutional boundaries. This problem is vexing, especially when dealing with archival collections of interest to humanities scholars. Such collections are topically narrow and deep. Archivists typically must implement their own specialized controlled vocabularies, because generalized systems (such as the Library of Congress Subject Headings, or LCSH) do not provide enough specificity. This problem has been encountered by groups ranging from the MetaScholar Initiative at Emory University to the UC San Diego Union Catalog of Art Images (UCSD UCAI) project. Automated mechanisms (such as the semantic clustering experiments described in the next section) for categorizing multi-institutional aggregations of metadata could potentially be applied ex post facto to metadata aggregations to remediate these problems.

- **Problems in Searching Across OAI and Web Realms:** The OAI-PMH is a nearly perfect mechanism for distributing metadata from databases and other dynamic content management systems. While the protocol is steadily increasing in deployment and availability, there is an enormous and still growing realm of Web content providers that are unlikely to soon expose their metadata via the protocol. This produces problems for services attempting to provide comprehensive search capabilities for some targeted subject domain. Specifically, a lot of the information that researchers would ideally like to be able to search is spread across the separated OAI and Web realms.

- **Problems in Coordinating Federated Digital Library Infrastructures:** The success of *lightweight protocols* (see glossary) like the OAI-PMH has accomplished two things: provider services based on such protocols have proliferated, and integrating services have struggled to find effective models and frameworks for attempting to amalgamate such distributed provider services into larger systems that work as a virtual whole. Digital libraries are still evolving at a rapid pace and will likely remain loose assemblies of distributed infrastructures for some time to come. Current consortial efforts to establish interoperable digital library federations like the AmericanSouth.Org system must proceed from the fundamental assumption that such federations will be loosely coupled. Tightly coupled systems with strong underlying programming infrastructures are not practical given the foreseeable level of coordination between digital library efforts and the rate of change we are experiencing. Libraries need more mechanisms like the OAI-PMH that can provide *abstraction interfaces* (see glossary) via lightweight protocols. This approach offers the promise of enabling interoperability among systems that will remain loosely coupled.

Plan of Work

The project will produce three broad deliverables, described below. Each deliverable will include 1) development of a working experimental infrastructure applied to either or both the AmericanSouth.Org and MetaArchive.Org scholarly portals, 2) assessment by means of multiple techniques, and 3) reporting results to the profession, both in conference presentations and in the professional literature.

A. Semantic Clustering Experiments

Summary: In this experiment, open source semantic clustering software will be applied to several collections of information aggregated from multiple institutions in order to categorize this information and make it browsable by researchers.

Background: Semantic clustering techniques appear to be a promising approach to remediate the problems of categorization described above in relation to harvested metadata. The most prominent and successful technique that has emerged in recent years for semantic clustering is the support vector machine (SVM) class of algorithms. SVM is the best currently known technique for automatically categorizing information, and is currently anticipated to be a powerful tool for automated organization of metadata. Another long-standing technique for semantic clustering is multidimensional scaling (abbreviated MDS). MDS provides a simple technique for graphically displaying the similarities and relationships of clustered information, as opposed to simply categorizing information for related item browsing. MDS is therefore graphically complementary to the results of SVM categorization.

Rationale: Applying the SVM and MDS techniques to a series of metadata and Web information aggregations will constitute a valuable experiment in organizing unstructured information for purposes of scholarly communication. The collections of information under consideration below are typical in that they lack effective overall classification categories, and therefore cannot be browsed by subject category.

Benefits: This is a practical experiment in that we hope to use the organized information resulting from these experiments in the AmericanSouth scholarly portal. The experiment further has broad applicability and therefore potential benefit to many other projects seeking to organize unstructured information for scholarly communication purposes. An example of such projects is the UCSD UCAI project mentioned previously. Emory intends to share information on this topic with the UCAI and similar projects for mutual benefit.

Details: This experiment will use open source semantic clustering software. Emory will conduct an evaluation of the many available SVM software tools (see http://www.support-vector.net/software.html), and select one or more for this experiment. MDS is a general statistical technique that is supported by many open source statistical environments (an example is the R language environment, see http://www.r-project.org). Emory will use SVM to categorize various combinations of information of scholarly interest in the study of the culture and history of the American South and make this information browsable. Because SVM is most effective when subject experts provide guidance and feedback to the system, Emory will employ the scholarly design team of the AmericanSouth project to train the SVM system to produce effective interdisciplinary categories of use to humanists. A system for MDS visualization of clustered information will be developed and applied to the results of these clustering experiments for graphical display and comprehension of results.

Specific experiments: Emory will undertake the following specific experiments with semantic clustering techniques applied to scholarly information.

1) AmericanSouth Metadata Clustering. The purpose of this experiment is to test whether or not SVM OSS can (with minimal guidance from experts) effectively categorize and make browsable all DC metadata harvested in AmericanSouth, for use by scholars researching the culture and history of the American South. The resulting body of winnowed and categorized metadata will be made browsable via the derived categories. Effectiveness of the categories will be gauged by feedback from scholarly consultants (see section on staff resources).

2) AmericanSouth Web Clustering. This experiment will test whether SVM OSS can categorize and make browsable the crawled Web content represented in the Web links section of the AmericanSouth portal (Web sites identified by scholarly consultants as including content of scholarly research value), tested under conditions similar to those listed in A1 above. The process and results of clustering Web content as opposed to metadata will be compared to understand similarities and differences.

3) Multi-Dimensional Visualization. A system will be developed to test whether effective means of visually displaying the SVM-derived subject categories is feasible using MDS graphical display of the results of both of the above experiments. Assuming that the experiment results in a display that provides comprehensible and worthwhile display of relationships, the following clustering results will also be visualized.

4) AmericanSouth Combined Clustering. This experiment will test if SVM OSS can categorize and make browsable a union set of harvested metadata and crawled Web content, to evaluate whether SVM semantic clustering can effectively organize such disparate information sets. This builds on the findings of MetaScholar Initiative projects to date, namely that both harvested metadata and crawled Web content should be integrated for comprehensive scholarly information discovery purposes.

5) American Memory Metadata Clustering. Finally, the project will conduct an experiment to test whether the DC metadata available from the American Memory OAI provider can be effectively culled and categorized for use by scholars researching the culture and history of the American South (as opposed to generalized American Studies). The resulting body of winnowed and categorized metadata will also be made browsable via the derived categories.

B. Experiments with Combined OAI / Web Search

Summary: Open source tools will be used to make a combination of harvested metadata and crawled Web content searchable in the context of the AmericanSouth portal.

Background: As mentioned, the MetaScholar Initiative has concluded that both harvested metadata and crawled Web content should be integrated for comprehensive scholarly information discovery purposes. However, this presents a challenge, as the two types of information are fundamentally different in nature, metadata being an abstraction of content, and Web pages being an instantiation of content. Both of the component tasks (harvesting/searching metadata, and crawling/searching Web content) are now relatively well understood. What is not well understood is the tasked of combined searching of these information realms.

Rationale: Experiments to bridge the OAI and Web information realms are needed by the MetaScholar Initiative, and would benefit other groups. There are two obvious ways that the OAI and Web realms might be bridged: subsuming OAI into the Web, or subsuming the Web into OAI. Each of these approaches should be evaluated.

Benefits: The findings of this experiment will have great practical benefit for the AmericanSouth portal, and will have potential application to virtually any other project seeking to automatically assemble a large amount of information for scholarly information discovery purposes. Emory intends to share information on this topic with other projects for mutual benefit.

Details: There are a variety of open source software tools that can usefully be tested for this purpose. The MetaScholar Initiative has accumulated extensive experience with the ARC software for OAI-PMH metadata harvesting and searching from Old Dominion University. Old Dominion plans to release a new, re-architected version of the software termed ARCHON that may include some capabilities for integrated metadata harvesting and Web crawling. There are a large number of open source Web search engines [Morgan, 2001] that could be adapted for this experiment. Finally, DP9 is a gateway service developed by Old Dominion University that enables indexing of an OAI data provider by an Internet search engine (see glossary for more information). DP9 is the logical mechanism to test the case of making the relevant OAI metadata searchable via the Web context. DP9 has not been tested by groups beyond Old Dominion to date.

Specific experiments: Two experiments will be undertaken in this area.

1) Combined Search Via Web Crawling. This experiment will test whether or not an open source Web search engine can be effectively applied to the union of the AmericanSouth harvested metadata (exposed via the DP9 gateway service) as well as the Web content represented by the AmericanSouth Weblinks. Focus groups of graduate researchers and scholarly consultants will evaluate the effectiveness of the resulting combined search system for scholarly discovery purposes.

2) Combined Search Via OAI-PMH. In this experiment, Emory will create an OAI provider for the metadata resulting from the experiment in clustering web content (# A2 above), and this metadata will be harvested and made searchable together with the existing metadata in AmericanSouth. Focus groups of graduate researchers and scholarly consultants will evaluate the effectiveness of the resulting combined search system for scholarly discovery purposes.

C. Federated Digital Library Framework Experiments

Summary: A framework for loosely-coupled federations of digital libraries will be iteratively developed as an improved mechanism for coordinating components of such an infrastructure.

Background: The success of lightweight protocols like the OAI-PMH has accomplished two things: provider services based on such protocols have proliferated, and integrating services have struggled to find effective models and frameworks for attempting to amalgamate such distributed provider services into larger systems that work as a virtual whole. There has been increasing attention to this problem in the research community. [Fox, 2002 and Castelli, 2002]

Rationale: The MetaScholar Initiative and other distributed/federated digital library infrastructures need better organizing frameworks for coordinating the operations of loosely coupled constituent systems, and enabling an extensible scheme for specifying proposed additions to such infrastructures. Experiments to utilize emerging industry standards such as the *Web services framework* (see glossary) and research standards such as the OAI-PMH would address this need.

Benefits: This experimental work will contribute to Emory's efforts to increase the usefulness of the internet for scholars, and potentially might have broader impacts on humanities scholarship if it works well. If the framework developed is flexible enough that various digital library services could modularly interact and share information then a large number of initiatives might stand to benefit. As a hypothetical example, if the Perseus Digital Library and AmericanSouth.Org could collaboratively build up interoperable Web services, both digital libraries would benefit.

Details: There are a number of promising new standards that will be utilized in this experiment. The OAI-PMH will be used as the underlying mechanism for distributing configuration and status information of virtual digital library systems. *Web Services Definition Language* (WSDL, see glossary) expressions will be disseminated via this OAI-PMH mechanism, representing the metadata for the digital library services of modular federations. The master configuration specifications for this framework will be expressed using the 5SL standard developed by Virginia Tech. [Fox, 2002]

Specific experiments: Emory will undertake two experiments:

1) Initial Federated Framework. An initial framework will be designed and implemented in the MetaArchive portal as a means of dynamically configuring virtual digital library federations. The only services that this initial framework will necessarily provide as modules are a federation coordinating service, an interface to a semantic clustering service, and a combined OAI/Web search service. The test to experimentally apply to this framework is whether or not it effectively enables the rapid and flexible creation of new federated digital libraries. Through this work, Emory seeks to develop a simple framework based on OAI that is both lightweight enough to be easily added to existing services and an effective means of configuring recombinant federations of digital library services.

2) Revised Federated Framework. Emory will design and implement a revised framework that will attempt to include targeted connection modules for a selection of other digital library services, such as the CLiMB toolkit from Columbia and the NITLE semantic indexing toolkit. The experimental test here is simply whether or not a working system can be devised incorporating these other tools in addition to the previous tool set. Through this work, Emory will explore the feasibility of a lightweight strategy for federating digital library services more generally, in the same way that the OAI-PMH enables simple federation of metadata resources. If this can be demonstrated, it will constitute a powerful approach for integrating digital library services across institutions.

Glossary & Acronyms

Abstraction Interfaces / Layers: A key concept that has emerged in software programming during previous decades is the concept of abstraction, or combining software routines with similar functions as one logical component (or layer) of an overall system. Interaction between such layered components is constrained through specified software interfaces to ensure that such layers are modular to the other parts of the overall system. In other words, they can be extensively modified internally without affecting the requirements for how other layered components interact with them. The concept of abstraction interfaces is the basis for many practices in software programming today, including programming libraries, software modules, and application programming interfaces (API's). Greater modularity in software applications enables more flexible development efforts, in which disparate teams can competitively work on separate interchangeable software modules that interoperate through well-defined programming interfaces. The OAI-PMH carried this concept to the metadata application level, decoupling metadata storage and searching functions in favor of a simple abstraction interface for sharing metadata.

DP9 Gateway Service: "DP9 is a gateway service that enables indexing of an OAI data provider by an Internet search engine. DP9 does this by providing a persistent URL for repository records, and converting this to an OAI query against the appropriate repository when the URL is requested. This allows search engines that do not support the OAI protocol to index the "deep Web" contained within OAI compliant repositories." (summary quoted from the DP9 Website at URL: http://arc.cs.odu.edu:8080/dp9/about.jsp)

Dublin Core (DC): This is the only form of metadata that the OAI-PMH mandates that providers supply, and consequently is the only form that can consistently be aggregated via harvesting. To date this has been synonymous with **Unqualified Dublin Core (UDC)**, a set of fifteen broad metadata elements. A major problem with UDC is the lack of prescriptive practice and standardization in use of these metadata elements. **Qualified Dublin Core (QDC)** has been under development for some time and offers the promise of more refined metadata.

Lightweight Protocols: A trend in the development of communication protocols in recent years has been to aim for simpler protocols that are consequently easier to implement than more fully featured protocols. An example of this trend is the success of LDAP (Lightweight Directory Access Protocol) as a simplification of the very complex X.500 protocol standard for networked directory access functions. Simplified (or lightweight) protocols are considerably easier to implement on top of existing communication systems, and therefore more likely to be implemented. The OAI-PMH is frequently cited as a lightweight protocol because it was explicitly designed to be easier to implement than other metadata dissemination standards such as Z39.50.

Multidimensional Scaling (MDS): A family of models in which multiple quantifiable aspects (or dimensions) of information items are represented geometrically to gain an understanding of comparative relationships between the items. MDS is a broadly applicable statistical technique, an has been used for decades in many fields ranging from market research to psychology to better understand how information with many facets is interrelated. [Romney et al. 1972] MDS applications often represent these relationships through geometric scatterplots of items, in which the many dimensions measured are simplified into two-dimensional images. MDS has many applications to the display of information. [Kruskal, 1978 and Borg, 1997] While MDS has occasionally been applied to bibliographic records [McGrath, 1983], it has not (to our knowledge) been applied to metadata aggregations harvested via the OAI-PMH.

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): An effort to develop and promote better mechanisms for sharing and using metadata, resulting from a 1999 meeting in Santa Fe, New Mexico. See http://www.openarchives.org.

Open Source Software (OSS): When the source code for particular software is openly available to the public, it is termed "open source". The typically means that the software is freely available and recombinable with other software, but this is not necessarily the case, as the creator may release the software under a variety of restrictions and licenses. There are many informative Websites on the open source movement; a representative example is http://www.opensource.org.

Support Vector Machines (SVM): This is a recently introduced technique for organization of information through statistical indexing of semantic content. SVM was first introduced by Vladimir Vapnik and his co-workers [Vapnik, 1992, 1995], and has been successfully adapted to applications in many fields such as bioinformatics, handwriting analysis, and text categorization [Joachims, 1998 and Cristianini, 2000]. The task of this class of algorithms is generally to detect and exploit complex patterns in data. SVM systems accomplish this by clustering, classifying, and ranking data by applying mathematical functions known as kernel methods to statistical properties of the data. SVM techniques are considered particularly accurate if they can iteratively be applied to a body of data with guiding feedback from subject specialists.

Semantic Clustering: This phrase refers to a broad approach to automatically organizing information by clustering similar items together via computerized algorithms. This topic has been investigated extensively during the past three decades of information science research. Several significant breakthroughs in this area have occurred during the last five years, notably the technique of *support vector machines*, a specific technique of semantic clustering. For a good overview of the overall trends in semantic clustering research, see Willett, 1988.

Web Services Definition Language (WSDL): "...an XML language for describing Web services. This specification defines the core language which can be used to describe Web services based on an abstract model of what the service offers." (from http://www.w3.org/TR/wsdl12)

Web Services Framework: The W3C and affiliated industry groups have in recent years endorsed an overarching scheme for interoperation of XML-based lightweight protocols for the Web, usually termed the Web services framework. The following text is from a W3C document on this framework (see http://www.w3.org/2001/03/WSWS-popa/paper51): "The XML Protocol work is the foundation for a Web Service framework within which automated, decentralized services can be defined, deployed, manipulated and evolved in an automated fashion. The purpose of this document is to outline a framework for evolving XML Protocol's functions. This framework provides a structure for integration and a foundation for protocols that will support the needs of such service-oriented applications. The goal is a scalable, layered architecture, one that can appropriately meet the needs of both simple and extremely robust high-volume deployments. As with other Web technologies, the focus is on enabling ubiquitous interconnectivity of entities and organizations dispersed throughout the world.

References

[Borg, 1997] Borg, Ingwer, and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications (Springer Series in Statistics).* Springer, 1997.

[Castelli, 2002] Castelli, Donatella, and Pasquale Pagano. “OpenDLib: A Digital Library Service System, in Research and Advanced Technology for Digital Libraries,” in the *Proceedings of the 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002.* Lecture Notes in Computer Science 2458, Springer, 2002.

[Cristianini, 2000] Cristianini, Nello, and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, 2000. URL: http://www.support-vector.net

[Fox, 2002] Goncalves, Marcos, and Edward Fox. “5SL - A Language for Declarative Specification and Generation of Digital Libraries,” in the *Second ACM/IEEE Joint Conference on Digital Libraries, Portland, Oregon, USA, July 14-18, 2002*. URL: http://www.dlib.vt.edu/projects/5S-Model/p117-goncalves.pdf

[Joachims, 1998] Joachims, Thorsten. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Proceedings of the European Conference on Machine Learning, Springer, 1998. URL: http://www.cs.cornell.edu/People/tj/publications/joachims_98a.pdf

[Kruskal, 1978] *Multidimensional Scaling.* Sage Publications, 1978.

[McGrath, 1983] McGrath, William, and Thomas Hickey. *Research Report Prepared for OCLC on Multidimensional Mapping of Libraries Based on Shared Holdings in the OCLC Online Union Catalog.* OCLC, 1983.

[Morgan, 2001] Morgan, Eric L. “Comparing Open Source Indexers.” *Infomotions Musings*, May 2001. URL: http://www.infomotions.com/musings/opensource-indexers

[Romney, 1972] Romney, A. Kimball, Roger Shepard, and Sara Nerlove. *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences.* Seminar Press, New York, 1972.

[Vapnik, 1992] Vapnik, Vladimir, Bernhard Boser, and Isabelle Guyon. “A Training Algorithm for Optimal Margin Classifiers.” Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory (COLT 1992), July 27-29, 1992, Pittsburgh, PA, USA. ACM 1992, pp. 144-152.

[Vapnik, 1995] Vapnik, Vladimir. *The Nature of Statistical Learning Theory*. Springer, 1995.

[Willet, 1988] Willett, Peter. “Recent Trends in Hierarchic Document Clustering: A Critical Review.” *Information Processing & Management,* Vol. 24, No. 5, pp. 577-597, 1988.