Findings from the Mellon Metadata Harvesting Initiative

Martin Halbert[1], Joanne Kaczmarek[2], and Kat Hagedorn[3]

[1] Woodruff Library, Emory University, Atlanta, GA 30322, USA
mhalber@emory.edu
[2] University of Illinois, 1408 W. Gregory Dr., Urbana, IL 61801, USA
jkaczmar@uiuc.edu
[3] Digital Library Production Service, University of Michigan, Ann Arbor, MI, 48109, USA
khage@umich.edu

Abstract. Findings are reported from four projects initiated through funding by the Andrew W. Mellon Foundation in 2001 to explore applications of metadata harvesting using the OAI-PMH. Metadata inconsistencies among providers have been encountered and strategies for normalization have been studied. Additional findings concerning harvesting are format conflicts, harvesting problems, provider system development, and questions regarding the entire cycle of metadata production, dissemination, and use (termed metadata gardening, rather than harvesting).

1 Introduction

The Open Archives Protocol for Metadata Harvesting (OAI-PMH) has seen rapid adoption by online repositories since the first release of the protocol in 2001. The protocol was designed to enable many kinds of services based on a relatively simple set of six metadata dissemination mechanisms. There has been widespread anticipation that the protocol would enable a new generation of novel information discovery services. [1]

The Andrew W. Mellon Foundation fostered the adoption and exploration of the OAI-PMH through seven project grant awards in 2001 totaling US$1.5M to the following intuitions: the University of Illinois at Urbana-Champaign (UIUC), the University of Michigan, Emory University, the Southeastern Library Network (SOLINET), the University of Virginia (UVA), the Woodrow Wilson International Center for Scholars (WWICS), and the Research Libraries Group (RLG). [2] The Mellon Foundation sought through these projects to enable practical research and experimentation into services and methods of utilizing the OAI-PMH in support of scholarly communication.

This paper describes research findings from four of the seven Mellon-funded projects, specifically those of UIUC, Michigan, Emory, and SOLINET. These four institutions have developed public discovery services that aggregate metadata from many institutions. The remaining three projects were focused on applications of the OAI-PMH internal to their organizations. Their findings will not be reported here.

This paper does not provide either instructions for implementing the protocol or extensive systems architecture reviews of the four projects described. Rather, the focus will be on major conclusions of these projects regarding the use of the OAI-PMH for discovery services of various kinds and directions for future research into applications of metadata harvesting.

The UIUC and Michigan projects have collaborated closely, and are reported first. The Emory and SOLINET projects were conjoined, and are reported jointly as the MetaScholar Initiative.

2 University of Illinois - Cultural Heritage Repository

The University of Illinois Library developed middleware tools for harvesting OAI-PMH–compliant metadata and built a Web portal through which cultural heritage resources could be discovered. This provided a basis for evaluating the potential utility of the OAI-PMH in search and discovery services, and helped identify issues that arise when implementing OAI-based services in this domain.

We developed both data provider and service provider open source tools in Visual Basic and Java which are available for download from the SourceForge Website. [3] These tools were subsequently used to create a searchable aggregation of metadata. [4] This aggregation represents a heterogeneous collection of cultural heritage materials that represent artifacts, digitized texts and images, video, sheet music, library holdings, and personal papers and manuscripts. We applied the software tools to the project in two ways. In the first instance we used the tools to provide "surrogate" data provider services for institutions interested in sharing resources but not quite prepared to provide these services themselves. In the second instance we acted as an aggregator using the service provider tools to harvest over 2 million records from across 39 institutional collections. Once harvested these records were then indexed and made searchable using the XPAT software developed and supported by the University of Michigan Library.

Our findings support the notion that the OAI-PMH is a reasonable structure for providing system interoperability among organizations or individuals interested in resource sharing and discovery. The tools are reasonably easy to install and maintain with modest ongoing support from technical staff. The service provider tools are capable of running multiple instances on one or more workstations. While the Java version of the service provider tools suffers from Java's known memory footprint issues, preliminary tests with the Visual Basic tools indicate the use of one workstation running five simultaneous harvest jobs could harvest up to ten million records in one twenty four hour period. Although the protocol seems to be a viable option for resource sharing, our Cultural Heritage Repository still encountered some distinct challenges.

2.1 Slow Adoption of the Protocol

Potential data providers were slow in adopting the protocol. Reasons for this slow adoption seem to be based more on available financial and personnel resources than

on a lack of interest in participation. This lagging participation was addressed by hosting data via "surrogate" data provider services for interested institutions. Our data hosting allowed us to collect batch dumps of metadata records, process them to conform with the OAI-PMH, and then provide server space for them from which they could be harvested by the project's harvesting service. We believe the slow adoption by potential data providers was foreseeable. At the start of the Mellon funded projects there was not yet enough momentum behind the OAI-PMH to illicit a greater commitment from potential data providers. As the Illinois-supported surrogate data provider services are phased out only sites we can directly harvest will remain in our Cultural Heritage Repository. Providing this temporary service allowed us to work on metadata inconsistency issues early on in the project.

2.2 Inconsistent Applications of Metadata

A second challenge, and one faced by all data aggregators, was deciding how to manage the inconsistent application of metadata applied to data provider records. These inconsistencies are compounded when aggregating metadata records from heterogeneous collections. Various data provider communities such as libraries, museums, archives and special collections, tend to use the required Dublin Core (DC) schema differently. The DC schema is flexible, easily understood and can be used to represent a variety of resources. However it is precisely this flexibility that adds to the inconsistencies found in the metadata records we have harvested. Metadata authoring practices indicate great variances in the levels of description used. Also, as all DC elements are optional in the OAI-PMH, there is no guarantee a particular element will be found across all records. Similarly, since all DC elements are repeatable within one record, the number of occurrences could have a significant impact on the outcome of results ranking or sorting. The Illinois project team used common normalization techniques, mapping terms such as 'photo', 'picture', and 'photograph', to 'image' in the Type element. [5] We also normalized on the Date element and temporal uses of the Coverage element and continue to explore techniques for providing more usable result sets to a person searching the repository.

2.3 Importance of Context

The context from which some metadata records are extracted can play a significant role in the value perceived by a person whose search results have revealed these materials. Maintaining the context associated with particular records proved to be a challenge for the Illinois project. An example of records that might not be as useful when taken out of context is found in the use of finding aids used by the archival and special collection communities. Finding aids typically describe a collection rather than an individual item within a collection. The Cultural Heritage Repository received 8,730 finding aids marked up with Encoded Archival Description (EAD) from data providers. EAD is widely used in the archival and museum communities to encode metadata representing a broad spectrum of materials ranging from manuscript collections to individual photos, letters, and artifacts. The project team decided to

expose as much granularity from the EAD files as possible. As such we developed algorithms to derive multiple DC metadata records, each describing an individual item, from each EAD encoded finding aid. [6] The process generated over 1,515,00 records from the original set of 8,730 finding aids. This level of granularity presented a particular challenge, as the types of records generated did not provide sufficient context for each record thereby leaving the possibility of a very unsatisfactory result set from a search. While searching discrete records embedded in an EAD finding aid may be desirable, to obtain maximum benefit, it is also imperative to provide a reference to the context of any one of these records. To this end, we developed a process that provides a link to the full EAD finding aid for each unique record, displaying the record embedded in its surrounding finding aid contextual framework. This display pops up in a separate browser window, conveniently providing a reference for the searcher to understand which archives or special collection record series generated the original record of interest. This strategy seems to make sense for displaying records aggregated from very different collections and marked up with very different metadata schemas.

2.4 Future Directions for Research at University of Illinois

Since the OAI-PMH allows for records with multiple metadata schemas to be harvested, providing they are minimally expressed in DC, it would not be unrealistic for service providers to harvest records in multiple formats. These records could then be served up to multiple end-user communities according to the most appropriate format for each particular group. As the OAI-PMH continues to mature, the work ahead for aggregators will focus on the delivery of services to end-users, packaging information on demand, and maintaining easy links to the contextual setting for records represented in heterogeneous repositories. The University of Illinois is engaged in some aspects of this work as it continues to extend the OAI-PMH on other grant-funded projects.

Through its National Leadership Grants (NLG) program, the Institute of Museum and Library Services (IMLS) is enabling the development of hundreds of significant new digital collections. In September 2002, IMLS awarded the University of Illinois a National Leadership Grant for a three-year research project to promote the visibility, adaptability, and interoperability of IMLS digital collections. The primary goals of the IMLS Digital Collections and Content project [7] are to:

- Create a registry of digital collections funded by the IMLS between 1998 and 2005
- Implement a search and discovery system for item-level content in these collections using the OAI-PMH.
- Research best practices for interoperability among diverse digital content and for supporting the interests of diverse user communities.

Funded by the NSF, the University of Illinois Library faculty are also working with faculty from the Department of Theoretical and Applied Mechanics (TAM) and researchers from Wolfram Research, Inc., to develop second-generation capabilities

for two digital libraries to support mathematics, engineering, physics, and applied sciences education using the OAI-PMH. Metadata has been harvested from Eric Weisstein's World of Mathematics (MathWorld) Website and Wolfram Research's Mathematical Functions Website. This material is then being exposed using the OAI-PMH for inclusion in the National Science Digital Library.

3 University of Michigan - OAIster

OAIster, at the University of Michigan, University Libraries, Digital Library Production Service (DLPS), was one of the Mellon grant-funded projects designed to test the feasibility of using OAI-PMH to harvest digital object metadata from multiple and varied digital object repositories and develop a service to allow end-users to access that metadata. We developed a system to harvest, store, and transform this metadata, using the UIUC Java harvester and our own Java transformation tool. We utilized the capabilities of Digital Library eXtension Service (DLXS) middleware to transform the metadata into a standard format (DLXS Bibliographic Class), build indexes and make the metadata searchable through an interface using the XPAT search engine. In addition, we tested the resulting interface in-house and remotely with users in two sets of sessions.

The unique feature of OAIster is that it provides access to metadata pointing to actual digital resources. Harvested metadata that has no corresponding digital resource is not indexed in OAIster. This ensures that users can access the resource itself. OAIster can be found online at http://www.oaister.org/, with over a million records available from over 140 institutions. Because of the large number of records indexed (and growing by more than 1000 records every month), OAIster is becoming a “one-stop shopping” interface to any digital resource users might need.

As a result of this year-long project, the University of Michigan project staff encountered a number of issues that should be taken into account for future uses of OAI metadata creation and harvesting.

3.1 Metadata Harvesting Issues

Scheduling harvesting of metadata is challenging, as long harvesting efforts often end up overlapping, and thus cause problems with memory-intensive, concurrent processes. Specifically with the Java harvester, we would at times receive out-of-memory errors, and a harvest would fail as a result. We were unable to implement an automated solution for harvesting, as the result of time concerns, so harvesting has remained a manual, time-intensive process.

We found it quite useful to have the ability to harvest a repository completely from scratch, in addition to on an incremental basis. Because of harvester and human hiccups, records could “go missing” on a variable basis. We developed a perl script that runs on the harvester database to clean out any administrative metadata about the harvested records of a repository, allowing us to re-harvest all records from that repository.

In over 10% of the repositories currently being harvested, we have encountered XML validation errors. Some data providers have not been strict in conforming to the UTF-8 encoding standard, and our harvester fails as a result when gathering records from these repositories. If we request that the harvester gather records under loose validation, we receive "junk" records as well as normal records, although in general we receive more records than if we don't harvest using loose validation.

3.2 Metadata Variations

Repositories vary widely in terms of the types of records they offer. They differ in formats (e.g., text, video), academic levels (e.g., graduate student theses, peer-reviewed articles), and topics (e.g., physics, religious studies), among others. And, the repositories vary significantly in the quality of their metadata, including their use of the Dublin Core (DC) encoding format. Although all repositories must at the least encode their data in DC before being OAI compliant, institutions use certain elements more frequently, ignore other elements completely, or include namespace declarations in non-standard places, e.g., included with a DC Title tag (`<title xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">`), which ultimately means we need to create many different kinds of XSL stylesheets as part of our transformation process.

Normalization of the DC Resource Type element was our attempt to standardize some of the metadata, so that users could search more effectively using this element. The transformation tool uses a normalization table, to transform DC Resource Type values such as "Book" and "paper" to the normalized value "text," and values such as "illustration" and "Picture" to the normalized value "image." The table was manually created from a retrieval of unique DC Resource Type values among all harvested records. The resulting normalized value is placed in a new Bibliographic Class element, specifically created for this purpose. Consequently, in the interface, users are then able to limit their search to those records in which the normalized element value equals, for example, "text." Admittedly, the method we used for normalization is not perfect, since each time a new repository, with potentially different varieties of values for DC Resource Type, is added to our service, the normalization table must be created anew. The normalization of metadata will eventually require the use of a thesaurus or controlled vocabulary, and automated methods for gathering the normalization table.

3.3 Duplicate Metadata Records

We have encountered duplication of harvested records in two ways. We have harvested nearly every OAI-enabled data provider repository, including those that are aggregators of repositories. Because of this, we have records in our system from original repositories and from aggregator providers collecting original repositories. As an example, a search performed in OAIster for "double-well Duffing oscillator" retrieves two records, exactly the same, but one was harvested from the arXiv.org Eprint Archive repository (an original repository) and one harvested from the

CiteBase repository (an aggregator). The decision of whether or not to harvest from aggregator repositories is made more complex because these aggregators contain records that are not currently available through OAI channels, and they do not always contain all the records of a particular original repository.

Aggregator providers can be very useful for service providers who are unable to harvest certain repositories, as was the case for us in our initial attempts to harvest the arXiv.org Eprint Archive repository. However, aggregators are somewhat problematic as they can be both service providers and data providers. Providing metadata that is also being provided by the original repository complicates the issue of duplicate records. Worse, if the original repository is also a service provider, there is the danger that they could (potentially unknowingly) harvest their own data. The use of the OAI Provenance element could alleviate these problems, however it's not currently widely used by data providers and most harvesters do not have the capability to utilize it.

3.4 Granularity of Resources

Granularity or specificity of digital objects will become more of an issue once more metadata is available. We foresee that it will confuse the user to be able to access separate records of the scanned artwork of the Peggy Guggenheim collection, but only be able to access a single record of a book of Emily Dickinson poetry and not records for each of the poems themselves. With so many harvested records, it is not possible to manually determine the specificity of each record, and mirror that in the interface appropriately (e.g., an hierarchical display). The future of the protocol may allow more automated approaches to solving this problem, especially if the OAI Set element is more fully realized and used. (The currently defined Set element allows a data provider to indicate which groupings a record belongs to, e.g., a "low temperature physics" set.)

3.5 Resource Restrictions

Rights and restrictions associated with metadata, and associated with the digital objects themselves, are also an issue. For instance, in our role as data provider, we provide metadata of our own collections that contain digital objects restricted to certain communities (e.g., CIC institutions), and metadata restricted due to contractual obligations with the originator. We don't OAI enable the latter, but we do enable the former, if available.

There are issues surrounding display of metadata pointing to restricted digital objects, not the least of which is that it may confuse users if they attempt to get to a digital object and are prevented from doing so. However, there are reasons for making this metadata available. Limiting access to records because the digital objects are restricted to a certain community ends up limiting that community's total access. At DLPS, we have digital objects that fit this pattern, and we make sure to include information in records that clarifies who is able to access those objects.

The large majority of metadata that we harvest displays the DC Rights element, with values supplied by the data provider. It would be beneficial to test differing

levels of restrictions and rights on digital objects with users to better answer how best to serve this metadata. Since rights information in the metadata varies widely, and currently there is no standardized method of indicating restricted digital objects (i.e., an OAI "yes/no" toggle element), testing this might inform future development of the protocol and data provider conformance.

3.6 Future Directions for OAIster

The University of Michigan intends to continue researching the use of OAI in a variety of ways. We expect to focus our efforts on:

- Determining a method for handling duplicate records.
- Normalizing more elements, such as DC Language.
- Providing high-level topical (or similar) browsing capabilities, perhaps drawing on the OAI Sets functionality.
- Working with UIUC on data mining research to offset issues related to metadata inconsistency.
- Targeting particular audiences within the research community.
- Collaborating with other projects that could benefit from using OAIster, e.g., giving researchers the ability to find digital objects while developing their courses online in a learning object environment.

4 Emory University and SOLINET – MetaScholar Initiative

Two of the projects funded by Mellon were based in Atlanta, Georgia. The MetaArchive project, proposed by Emory University, seeks to explore how information about targeted subject collections held in small-to-medium sized library archives could be most effectively disseminated through OAI-PMH metadata harvesting services. The AmericanSouth portal was the second Atlanta project, and was proposed by the Association of Southeastern Research Libraries (ASERL), a component of the SOLINET non-profit organization. AmericanSouth.Org is intended to be a scholarly portal for the study of the culture and history of the American South. These two projects were conjoined in a unified effort termed the MetaScholar Initiative in order to take advantage of complementary aspects of the two projects. Both projects are focused on cultural research materials associated primarily with regional subject topics. The projects have produced three portals, accessible at http://AmericanSouth.Org, http://MetaArchive.Org, and http://MetaScholar.Org (containing information about the overall initiative).

During 2002 the MetaScholar Initiative installed OAI data providers at partner institutions and established a central portal infrastructure to harvest and index metadata, as well as manage contributed contextual content pieces.

MetaArchive Project. MetaArchive project programmers worked with archivists and technologists at institutions and archival operations of many different types and scales. A special project focus has been on smaller college library archives, but we have also worked with large research libraries, church record repositories, university

data centers, and a museum. Several dozen OAI data providers were installed during the course of the project, representing multiple archival collections.

A difference between the philosophies of the MetaScholar Initiative and many other OAI harvesting projects is that metadata for both print and online materials have been aggregated. This is because the focus of our projects is raising visibility of scholarly resources generally, regardless of access mechanisms. We are interested in studying the processes whereby institutions of multiple categories can utilize the OAI-PMH in support of information dissemination, and especially how centralized common-good technical infrastructures can be deployed to benefit smaller, underfunded archives. Archives are being selected using a collection development model, with roughly 30K records in the current MetaArchive databases. We wish to keep the MetaArchive databases small and targeted to the two subject domains of the project, papers of selected Southern political figures and religious institutions.

AmericanSouth Project. The AmericanSouth project has focused on a more homogenous group of nine large ASERL research libraries and their archival holdings. The majority of these archival collections are digitized materials, but online access to items cited by metadata is not a harvesting requirement for AmericanSouth any more than in the MetaArchive project. The main criteria for harvesting are scholarly value and subject domain relevance of collections as judged by a team of scholars representing different disciplines and backgrounds. This scholarly design team also provides guidance on user interface features, as well as original online content in the form of online articles and authoritative contextual subject guides.

During 2002 MetaScholar Initiative programmers collaborated with researchers at Virginia Tech to install elements of the Virginia Tech Open Digital Library software [8] at AmericanSouth and MetaArchive partner sites in order to establish OAI data provider. The MetaScholar Initiative redeployed the ARC software [9] from Old Dominion and enhanced it with additional annotation and portal features to create two experimental portals, AmericanSouth.Org and MetaArchive.Org, that combined metadata harvesting and scholarly publishing capabilities.

The MetaScholar Initiative has independently encountered and can reiterate the all the findings reported by the UIUC and Michigan projects. Without repeating their points, there are some further elaborations that MetaScholar can offer, as follows.

4.1 Barriers to Adopting the Protocol.

THE OAI-PMH was not used, implemented, or adopted by the vast majority of libraries and archives in 2002, nor is this likely to change in 2003. This is because of a variety of barriers to adopting the protocol. This finding may appear inconsistent with the growth in numbers of OAI data providers and the simple implementation requirements of the protocol, but is nevertheless valid and significant. It is true that the numbers of OAI data providers are growing at a virtually geometric rate. [10] But despite this rapid growth in providers, the overall penetration of the protocol into the deep infrastructure of scholarly archives is still minor. Virtually no MetaScholar partners were planning on implementing the OAI-PMH before the collaboration began. Nor would the majority have been able to implement the protocol without the specialized programming assistance offered by the MetaScholar projects.

Collaborating institutions nevertheless acknowledge the clear benefits the protocol offers by enabling new means of information dissemination. The lack of priority in adopting the protocol is mainly related to the generally poor funding situation that research institutions find themselves in today, in which not only new technological initiatives, but almost any new initiatives are difficult to undertake. Where funding for new technological initiatives is available, it is primarily being expended on other fundamental infrastructural changes, mainly migrating data from obsolete or discontinued software systems (examples: mainframes, and PCs using the DOS operating system) to current digital library systems (examples: dynamic XML content management systems). These migrations are so encompassing that the technical staff available to libraries and archives are wholly occupied, and not alert to the opportunities presented by the OAI-PMH, at least in the 2002-2003 timeframe.

This finding leads us to strongly conclude that additional efforts to foster the adoption of the protocol in the library and archival communities are needed, and should be underwritten as special programs by central state and foundation funding agencies. Clearer metrics of the benefits of adopting the protocol are also needed in order to justify these expenditures.

4.2 Problems Associated with the Dublin Core Metadata Standard

Many of the problems reported by UIUC and Michigan (inconsistent applications of metadata, metadata variations, and granularity of resources) have also been difficulties encountered by the MetaScholar Initiative. We feel that these issues all stem inevitably from two facts: a) the primitive quality of metadata represented in the unqualified Dublin Core (UDC) format, and b) the fact that UDC is the only metadata format required by the OAI-PMH. These two facts are together problematic for services attempting to develop discovery services based on harvesting metadata using the OAI-PMH. While there is great discovery utility in being able to rapidly and easily harvest metadata from large numbers of distributed, heterogeneous systems, the search functionality in such discovery systems is compromised by unavoidable inconsistency in the metadata provided by varied sources.

Metadata Format Collisions. In addition to the problems identified by UIUC and Michigan, we have become deeply troubled by what we see as intractable conflicts or collisions between the approaches of groups using different metadata formats. Because MetaScholar has worked extensively with archives, we have particularly studied the problems that occur when metadata originally expressed in EAD is converted into UDC. There is a Procrustean dilemma that frequently occurs in these situations. The EAD format is primarily aimed at encoding a single finding aid document characterizing an entire collection, typically with series-level information but not usually item-level details. When converting finding aids into UDC records, the metadata wrangler in charge of the operation is often faced with two distasteful options: 1) creating a single UDC record for the entire collection (this single record must often discard much of the detail of the original finding aid in order to be of reasonable length for screen display), or 2) segmenting the series-level information into separate UDC records (that are often decontextualized by this segmentation). The UIUC and MetaScholar projects have sought middle-ground strategies in

segmenting large finding aids while trying to preserve context through hyperlinks, with varying degrees of success. While this problem has been prominent initially in the case of EAD/UDC metadata collisions, we believe that similar kinds of deep metadata conflicts will likely arise as other metadata formats are increasingly harvested into discovery services. Better guidelines for consistent use of differing metadata formats are needed.

4.3 Metadata Gardening

While services based on metadata harvesting have been explored by several projects, we believe there is still inadequate understanding of the entire cycle of metadata production, dissemination, reprocessing, and extended uses. The MetaScholar Initiative has conceptualized this cycle as "metadata gardening" in its work with participating institutions and scholars. We think that several of the issues identified by Michigan (which MetaScholar has also encountered) including duplication of records, confusing rights restrictions, harvesting problems, and metadata variations are all examples of issues that arise when the entire cycle is not coordinated. Additional problems in this area that MetaScholar has identified are uneven representation of topics, problems in browsing across collections, and poor understanding of the value that metadata harvesting services add to scholarly communication (see below).

Without better perspective on how to coordinate the entire cycle of cultivating metadata providers, harvesting, and finally organizing metadata for discovery, we feel that these problems will continue to compromise the effectiveness of services based on metadata harvesting.

4.4 Interactions with Scholarly Communication

One of the primary reasons that the Mellon Foundation funded these metadata harvesting projects was to explore how the OAI-PMH could benefit scholarly communication. The MetaScholar Initiative feels that there has been relatively little direct examination of this question in any of the Mellon projects that were undertaken. A basic conviction of the MetaScholar Initiative is that simple aggregations of metadata as such are not sufficient; an associated authoritative context of some user community is always needed to make a metadata discovery service useful. Trying to understand the specific value and opportunities that may be realized by close connections of metadata services with scholarly content is an issue that the AmericanSouth scholarly design team is taking up in 2003.

4.5 Relationships to Other Discovery Services

There is no consensus on the relationship of metadata harvesting services with traditional library tools such as the online catalog, or newly emergent tools such as Google and other Web search engines. This issue has dismayed information service

providers such as librarians at the campus level who now have yet another "one stop shopping" service to try to coherently articulate to students and other researchers. We are interested in answering the question of whether or not there are logical ways of creating federated search systems that can simultaneously query both OAI metadata aggregations as well as web search engines.

4.6 Future Directions for Research at Emory University and SOLINET

The MetaScholar Initiative will undertake the following projects in coming months:

- Fostering adoption of the OAI-PMH through workshops
- Examining metadata format collisions more closely
- Modeling the metadata gardening cycle
- Studying the benefits and interactions of metadata harvesting and scholarly communication
- Clarifying the relationship of metadata harvesting services and other discovery services through future testbed projects to explore the combination of metadata harvesting and web crawling

References

1. Lynch, C.: Metadata Harvesting and the Open Archives Initiative. ARL Bimonthly Report, Issue 217 (2001) (http://www.arl.org/newsltr/217/mhp.html)
2. Waters, D.: The Metadata Harvesting initiative of the Mellon Foundation. ARL Bimonthly Report, Issue 217 (2001) (http://www.arl.org/newsltr/217/waters.html)
3. SourceForge Website. (http://sourceforge.net/projects/uilib-oai/).
4. Illinois Cultural Heritage Repository Website. (http://oai.grainger.uiuc.edu).
5. Cole, T.W., Kaczmarek, J., Marty, P.F., Prom, C.J., Sandore, B., & Shreeves, S.L.: Now that we've found the 'hidden web' what can we do with it? The Illinois Open Archives Initiative Metadata Harvesting experience. In: Bearman, D., Trant, J. (eds): Museums and the Web 2002: selected papers from an international conference, (2002) 63-72. (http://www.archimuse.com/mw2002/papers/cole/cole.html)
6. Prom, C.J., Habing T.G.: Using the Open Archives Initiative Protocols with EAD. In: Marchionini G., Hersch, W. (eds): JCDL 2002: Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries (2002)171-180.
7. IMLS Digital Collections and Content project (http://imlsdcc.grainger.uiuc.edu)
8. Open Digital Libraries at Virginia Tech (http://oai.dlib.vt.edu/odl)
9. ARC SourceForge site (http://oaiarc.sourceforge.net)
10. Van de Sompel, H., Lagoze, C.: Notes from the Interoperability Front: A Progress Report on the Open Archives Initiative. In: Agosti, M., Thanos, C.(eds): Research and Advanced Technology for Digital Libraries: 6th European Conference. Lecture Notes in Computer Science, Vol. 2458. Springer-Verlag, Berlin Heidelberg New York (2002) 150.