The Open Archives Initiative: Realizing Simple and Effective Digital Library Interoperability

Hussein Suleman
hussein@vt.edu
Department of Computer Science
Virginia Tech, Blacksburg, VA, USA
+1 540 231-3615

Edward Fox
fox@vt.edu
Department of Computer Science
Virginia Tech, Blacksburg, VA, USA
+1 540 231-5113

Abstract

The Open Archives Initiative (OAI) is dedicated to solving problems of digital library interoperability. Its focus has been on defining simple protocols, most recently for the exchange of metadata from archives. The OAI evolved out of a need to increase access to scholarly publications by supporting the creation of interoperable digital libraries. As a first step towards such interoperability, a metadata harvesting protocol was developed to support the streaming of metadata from one repository to another, ultimately to a provider of user services such as browsing, searching, or annotation. This article provides an overview of the mission, philosophy, and technical framework of the OAI.

Keywords

Interoperability, harvesting, metadata, protocol, repository.

Biographical Information

Dr. Edward A. Fox holds a Ph.D. and M.S. in Computer Science from Cornell University, and a B.S. from M.I.T. Since 1983 he has been at Virginia Polytechnic Institute and State University (VPI&SU or Virginia Tech), where he serves as Professor of Computer Science. He directs the

Internet Technology Innovation Center at Virginia Tech, Digital Library Research Laboratory, Networked Digital Library of Theses and Dissertations, Curriculum Resources in Interactive Multimedia, and a number of other research and development projects. He has and continues to serve in various editorial and leadership roles, most recently including the co-editorship of the ACM Journal of Educational Resources in Computing and the general chair for the ACM/IEEE Joint Conference on Digital Libraries '2001.

Hussein Suleman is a PhD student working with Edward Fox at Virginia Tech. His research focus is on topics closely related to matters of interoperability. He served as part of the technical working group that produced the latest revision of the Open Archives Metadata Harvesting Protocol. In this capacity he implemented the standards on various platforms and also developed and actively maintains the Repository Explorer software that is used by the Open Archives community to rigorously test archives for compliance with the standards.

The Open Archives Initiative: Realizing Simple and Effective Digital Library Interoperability

Abstract

The Open Archives Initiative (OAI) is dedicated to solving problems of digital library interoperability. Its focus has been on defining simple protocols, most recently for the exchange of metadata from archives. The OAI evolved out of a need to increase access to scholarly publications by supporting the creation of interoperable digital libraries. As a first step towards such interoperability, a metadata harvesting protocol was developed to support the streaming of metadata from one repository to another, ultimately to a provider of user services such as browsing, searching, or annotation. This article provides an overview of the mission, philosophy, and technical framework of the OAI.

1 Introduction to the OAI

1.1 Historical Background and Context

The World Wide Web (WWW) is frequently thought of as the technology that revolutionized computer networking by effectively breaking down the barriers between the providers of content and the users of that content. The underlying idea was not particularly a novel one since the hypertext community has been investigating such avenues for decades. However, it was backed up by free, easy to utilize software that satisfied a need in the rapidly advancing networked community, and so it was immensely successful.

The WWW broke down a major barrier in making information freely accessible, but it also created information management problems for which simple solutions did not exist. One such problem is that of persistence: how can we guarantee that a digital object on the WWW will always exist? Another question has to do with authority: how much trust can we place in the

authenticity of a source of digital objects? These and other concerns led some individuals and organizations to begin creating managed repositories of digital information, nowadays called Digital Libraries (DLs), with additional and specialized services to enhance the users' experience beyond what the WWW had to offer.

While the WWW thrived because of its distributed nature, most DLs tried to provide one-stop shopping for users in specific communities. As the number of DLs increased, users looking for resources found that they needed to search through many DLs before finding what they needed. Most DLs are driven by databases; thus the popular search engines do not index their contents. As a result, search engines are not of much use to users who want to perform searches across multiple DLs.

In order to address this need, different approaches were taken by various communities of users. The Z39.50 (ANSI/NISO, 1995) protocol was designed for client/server access and adapted to federated searching, whereby a system performing a search operation on multiple repositories could send the query to all of them in a standardized format and then process the returned results as appropriate. The Harvest system (Bowman et al., 1995) attempted to gather metadata from websites and create a central searchable index. The Dienst protocol from Cornell University (Davis and Lagoze, 2000) and the STARTS protocol from Stanford University (Gravano et al., 1997) both implemented variations of federated search algorithms, where queries are sent to remote sites in real-time. Kahn and Wilensky's Repository Access Protocol (Kahn and Wilensky, 1995) allowed remote access to the contents of a repository, thus facilitating search and browsing operations. These projects had varying degrees of success, in most cases limited to large or research DLs where there was a commitment to building interoperability into the

systems. Smaller DLs were not prepared to make the investment in a complex protocol for interoperability, especially since the rewards were not immediately tangible.

In October 1999, a meeting of representatives of various existing archives was held in Santa Fe, New Mexico, USA, to address the concern that interoperability was beyond the reach of most DL systems. Delegates at this meeting included representatives of the Association of Research Libraries, Coalition for Networked Information, Council on Library and Information Resources, Digital Library Federation, Library of Congress, Networked Digital Library of Theses and Dissertations (NDLTD), Scholarly Publishing and Academic Resources Coalition, and various universities and research institutes. The primary focus of delegates was on facilitating the creation of a Universal Preprint Archive (van de Sompel et al., 2000) – a DL that contained all electronic pre-prints such as papers, articles, and theses. The result of this meeting, the Santa Fe Convention (van de Sompel and Lagoze, 2000), was an agreement among the parties to subscribe to a common standard for interoperability based on transfer of metadata from repositories using a minimal protocol and leveraging existing technology to achieve this.

1.2 Initial Technical Efforts

The Santa Fe Convention laid the groundwork for future efforts by defining the guiding principles of the Open Archives Initiative (OAI) (OAI, 2001) – principles that are largely unchanged after 18 months of further discussion within an expanding community of digital librarians and users of information.

At the Santa Fe meeting, it was decided that archives would be able to exchange metadata with one another using a modified subset of the Dienst protocol. As is often the case, however, this first iteration of the interoperability protocol led to much debate over semantics and ambiguities inherent within the specifications. Early implementations for the Computer Science Teaching

Center (CSTC, 2001) and the Physics Preprint Archive (arXiv, 2001) were based on subtly different interpretations of the protocol. Discussions among implementers of the protocol convinced some proponents of the Santa Fe Convention that more work was needed to make the protocol specification robust and thus truly standardized. This notion was formalized at two workshops and a technical committee meeting, which, along with a Steering Committee, guided the evolution of that initial protocol into its current incarnation.

1.3 Evaluation: Community and Technical Meetings

The OAI held two workshops in conjunction with the ACM DL2000 (San Antonio, USA, June 2000) and ECDL 2000 (Lisbon, Portugal, September 2000) conferences, where the initial work was evaluated and a future course was charted for the OAI.

Unlike the inaugural meeting, these workshops were openly advertised to digital library practitioners and they drew a broad range of participants from sectors of the community ranging from publishers to researchers. It was unanimously agreed that the initial protocol needed revision and that the OAI needed to broaden its scope to serve communities beyond its initial mandate of pre-print archives. To address these issues, a technical committee was formed and tasked with revising the protocol to eliminate the shortcomings that were recognized and to meet the needs of the larger OAI community. This committee met in September in Ithaca, NY, USA to launch an intensive period of writing, implementing and testing, which culminated in the official release of the OAI Metadata Harvesting Protocol in January 2001 (Lagoze and van de Sompel, 2001). This protocol, having undergone extensive alpha testing prior to release, promises to provide a simple mechanism for DLs to interoperate effectively.

2 Basic OA Concepts

2.1 Repositories and Open Archives

The words “Open Archive” frequently conjure up images of information access without any associated cost or restriction. While this is a goal for many proponents of the OAI, it would place too many restrictions on DLs that wanted to conform to OAI standards. So, the OAI defines an Open Archive (OA) simply as being an archive that implements the OAI Metadata Harvesting Protocol, thus allowing remote archives to access its metadata using an “open” standard.

A “Repository” is frequently used as a synonym for an OA. In the traditional DL context, a repository is a collection of digital objects, but in the context of the OAI, it has to be network accessible and it has to support the OAI Metadata Harvesting Protocol.

2.2 Harvesting and Federation

The first crucial decision made by the OAI was the selection of a method to achieve basic interoperability among repositories, with special emphasis placed on the ability to do cross-archival searching. It is generally considered that there are two major approaches to accomplish this: harvesting and federation.

Federation refers to the case where the DL sends the search criteria to multiple remote repositories and the results are gathered, combined, and presented to the user. Harvesting is when the DL collects metadata from remote repositories, stores it locally and then performs searches on the local copy of the metadata. Figure 1 illustrates the differences in data flow.

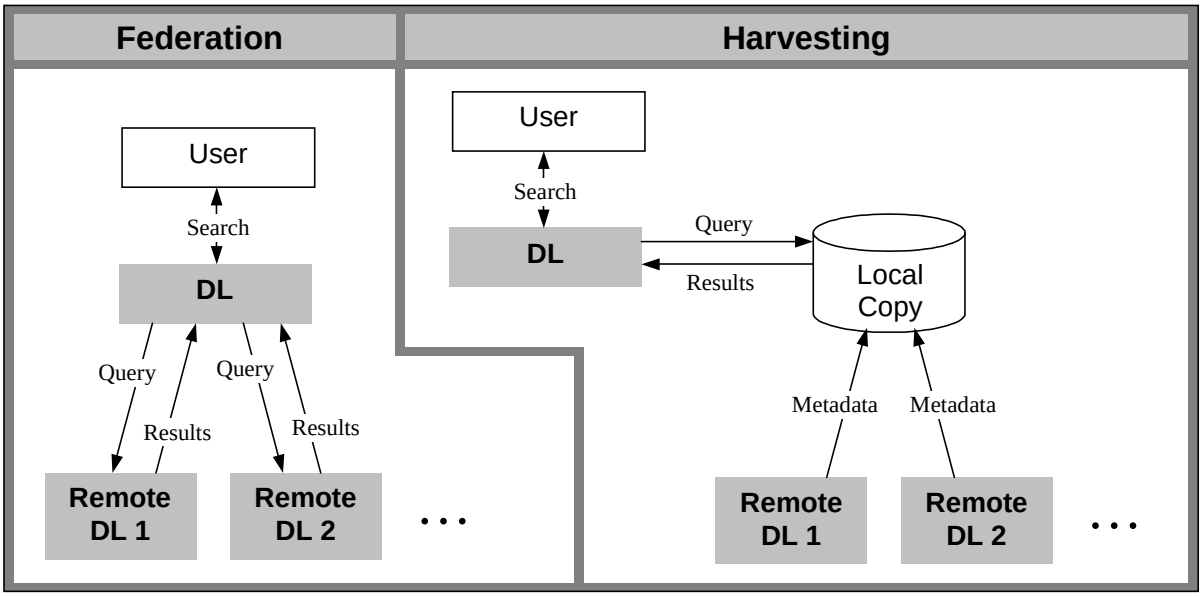

Figure 1. Data flow for federation and harvesting

Federation is a more expensive mode of operation in terms of network and search system constraints since each repository has to support a complex search language and fast real-time responses to queries. Harvesting requires only that individual archives be able to transfer metadata to the central DL. The frequency of queries, quantity of metadata, and availability of network resources also factor into this comparison but, in general, federation places a greater burden on the remote sites while harvesting reduces the demand on remote sites and concentrates the processing at the central DL site. Since it is more likely that providers of services, such as search engines, will expend the effort to store, index, classify, and otherwise manage searchable metadata, the OAI opted for harvesting, primarily as a means of lowering the barrier to interoperability for providers of data.

2.3 Metadata and Data

The question of what to harvest has been a contentious issue for many, as it is not obvious whether an archive should be sharing its metadata, its digital objects, or both. There are

advantages to exchanging complete digital objects since that would support operations like full-text search of text documents. However, in most instances DLs need only harvest metadata in order to provide search, classification, and related services. This approach was adopted by the OAI, with the implicit understanding that the metadata should contain pointers to the concrete rendition of digital objects.

2.4 Data and Service Providers

A data provider maintains a repository that allows external online access to its metadata through the OAI Metadata Harvesting Protocol. In the interest of brevity, "data provider" is sometimes used to refer to such repositories. A service provider is an entity that harvests metadata from data providers in order to present users with higher-level services. This distinction allows for a clean separation between the provider of data and the provider of services (as illustrated in Figure 2). This helps eliminate the current barrier to quality services that arose because of the historical connection between ownership of data and provision of services. In general, archives with large quantities of content prioritize information management over the provision of user services. On the other hand, if information management is not a primary function of an archive, more effort can be devoted to service provision. The OAI attempts to clarify and separate these approaches to present users with the best of both worlds.

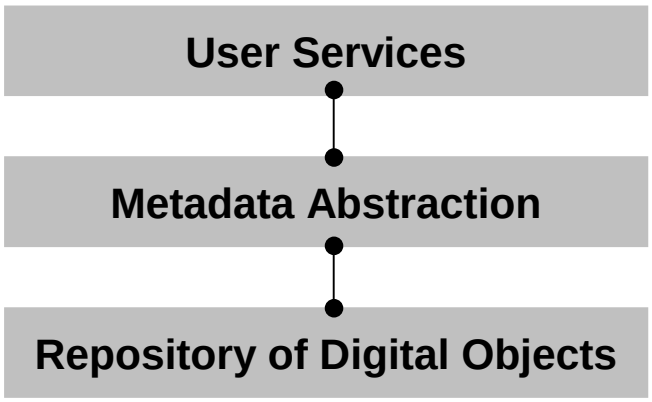

Figure 2. Layered organization of data storage and service provision

3 Technical Framework

3.1 Underlying Technology and Standards

3.1.1 HTTP

In creating a protocol for interoperability, it was considered prudent to build upon the existing infrastructure provided by the WWW. Thus, the OAI Metadata Harvesting Protocol is based on HTTP (Fielding et al., 1999), closely following the model upon which HTTP is based, and leveraging its mechanisms for redirection, error handling, and parameter passing. The Metadata Harvesting Protocol is a request-response protocol – the client makes requests for data and the server returns corresponding responses.

3.1.2 XML

While all requests are encoded as HTTP GET or PUT operations, responses are in XML (Bray et al., 2000) so as to allow for structure within the response data. This is especially well suited to handling the case where a service provider requests structured metadata from a data repository.

The frequently thorny issue of character encoding also has been deftly avoided by utilizing the support for such features in XML.

3.1.3 XSD and Namespaces

Data quality and correctness of implementations are crucial to the success of any new standard. To maintain such quality, automatic and manual testing can be performed on data providers to ensure conformance to the protocol. In both instances, this testing is largely driven by precise definitions of valid XML responses in the form of XML Schema Descriptions (XSD) (Fallside, 2000). While XSD is still a very young technology, it greatly enhances the ability to specify what constitutes a valid XML document. Service providers and conformance testing tools like the Repository Explorer (Suleman, 2001) use XSD tools to automatically validate XML responses from data providers.

XML tags may be grouped together by using a prefix for each group called a namespace. Namespaces are used to support the reuse of existing semantics and schemata, making validation a modular process. For example, some responses contain metadata fields embedded within a larger structure – in these cases, the metadata will use one namespace and the rest of the XML could belong to another namespace.

```
<testxml xmlns="space1" xsi:schemaLocation="space1 space1.xsd">

  <name>Joe Smith</name>
  <comment>testxml, name and comment are in the namespace space1</comment>

  <metadata xmlns="space2" xsi:schemaLocation="space2 space2.xsd">
    <date>2000-02-28</date>
    <description>
      metadata, data and description are in the namespace space2
    </description>
  </metadata>

</testxml>
```

Figure 3. Fragment of XML illustrating namespaces and schema locations

Figure 3 is a fragment of typical XML where namespaces are used to delineate tags from different namespaces by means of "xmlns" attributes. At the same time, the schema for each namespace is indicated with an "xsi:schemaLocation" attribute that creates a mapping from the namespace to the XSD document that can be used to validate the XML.

3.1.4 Dublin Core

It is compulsory that all open archives be able to generate metadata for all resources in unqualified Dublin Core (DC) (Dublin Core Metadata Initiative, 1997). This will ensure that service providers who do not understand any other metadata format will at least be able to glean the basic information about resources from their DC renditions. Dublin Core is almost never the best choice for metadata for any given repository, but its generality makes it suitable for interoperability in the context of the OAI and its application to various different types of repositories such as papers, theses, and multimedia documents. In addition to DC, repositories also may support other optional metadata formats that are better suited to represent the objects they contain. Thus, repositories connected with NDLTD also should support MARC or a newly devised thesis metadata standard (Atkins et al., 2001).

3.2 Sets

Sets are a special construct which allow a repository to expose its internal structure to service providers. It is not compulsory for an archive to support set constructs but it provides one more mechanism for selective harvesting. There are no predefined semantics for what constitutes a set so any use of sets must be by explicit agreement between data providers and service providers. For example, in the context of NDLTD, a national archive might have sets for each region, and subsets for each university.

3.3 Records

A record is the metadata bundle that is associated with a unique identifier. Usually, records correspond to simple digital objects but this is not necessary – records also could refer to collections or sub-objects. Records are encapsulated within a special structure that includes both the metadata and a header containing special fields used to support the harvesting operation. Figure 4 displays a typical record.

```
<record>
 <header>
  <identifier>oai:arXiv:alg-geom/9202004</identifier>
  <datestamp>1992-02-10</datestamp>
 </header>
 <metadata>
  <oai_dc xmlns="http://purl.org/dc/elements/1.1/">
   <title>Mirror symmetry and rational curves on quintic threefolds: a guide
          for mathematicians</title>
   <creator>Morrison, David R.</creator>
   <subject>Algebraic Geometry</subject>
   <description> We give a mathematical account of a recent string theory
                calculation which predicts the number of rational curves on
                the generic quintic threefold.</description>
   <date>1992-02-10</date>
   <type>e-print</type>
   <identifier>http://arXiv.org/abs/alg-geom/9202004</identifier>
  </oai_dc>
 </metadata>
</record>
```

Figure 4. Sample record from the arXiv open archive

3.4 OAI Metadata Harvesting Protocol

The OAI Metadata Harvesting Protocol supports 6 service requests that may be made to a repository. The protocol specifies the formats for HTTP queries and XML responses. These service requests are as follows: GetRecord, Identify, ListIdentifiers, ListMetadataFormats, ListRecords, and ListSets.

- GetRecord retrieves the metadata for a single object in a specified metadata format.
- Identify is a request for information about the repository as a whole. Returned is such information as the name of the repository, the version of the protocol, and the email address of the administrator. There also is an extension mechanism for a repository to specify additional information by supplying its own schema.
- ListIdentifiers lists identifiers for all objects or, if specified, those within a given date range and/or within a given set.
- ListMetadataFormats will return the list of all metadata formats supported by the archive, or all the metadata formats in which a particular object may be rendered.
- ListRecords lists complete metadata for all objects or, if specified, within a given date range and/or within a given set.
- ListSets lists the sets (and subsets, recursively) contained within the repository.

3.5 Flow Control

In principle, the OAI subscribes to the philosophy that the act of a service provider harvesting a repository ought not to interfere with the regular use of the archive by users through, for example, an existing WWW-based search and retrieval interface. However, some service requests have the ability to return very long response sets, e.g., ListContents, so to prevent

overloading the data provider can break result sets into chunks and return one chunk per request with a token being passed to keep track of the state of the system. Other flow control mechanisms like the ability to redirect a request or the ability to postpone a request are inherited from the underlying HTTP protocol.

3.6 Registration Services

Registration of conformant repositories is useful within communities with shared interests. For example, NDLTD will have a listing of all its member institutions that implement the OAI protocol. Registration can be automated by using the Identify service request to return information about an archive. On a more global scale, the OAI is attempting to register all repositories in order to provide a name resolution service from identifiers to repositories.

3.7 Expansion and Customization

The protocol has optional features in some strategic places to allow for future expansion. Most importantly, there is no restriction on which metadata formats may be supported as long as each one has an associated schema description. Also, the data returned by the Identify request includes optional sections for descriptions that conform to external schemata. Similarly, each record has an optional “about” section that may contain information about the metadata object, as opposed to the digital object associated with the metadata. Figure 5 displays a minimal metadata record with this optional section.

```
<record>
 <header>
  <identifier>oai:arXiv:alg-geom/9202004</identifier>
  <datestamp>1992-02-10</datestamp>
 </header>
 <metadata>
  <oai_dc xmlns="http://purl.org/dc/elements/1.1/">
   <title>Mirror symmetry and rational curves on quintic threefolds: a guide
          for mathematicians</title>
   <creator>Morrison, David R.</creator>
  </oai_dc>
 </metadata>
 <about>
  <oai_dc xmlns="http://purl.org/dc/elements/1.1/">
   <creator>University Library Cataloguing Service</creator>
  </oai_dc>
 </about>
</record>
```

Figure 5. Minimal metadata record from arXiv with optional "about" section

4 Requirements to be a Provider

4.1 Data Provider

Any archive that wishes to become a Data Provider must satisfy a few basic requirements. Firstly, and most importantly, the archive must have an online interface and a web server that can be used for the purposes of the protocol. Then, each record in the archive must be persistent or at least must contain a persistent identifier, each of which must be unique within the archive. It also is highly recommended that each archive have a unique archive name embedded within its identifiers for records so that OAI records can be globally unique – the OAI protocol suggests that unique identifiers adopt the form “oai:archive_id:record_id”. Finally, every record must have an associated date stamp to allow for harvesting of records within a particular date range.

4.2 Service Provider

Service providers may use the data they harvest as they wish to, within the boundaries laid out by the data providers. While the protocol does allow for an entire archive’s contents to be

harvested, it is expected that service providers will use date ranges to incrementally harvest new additions to a repository. This is illustrated in Figure 6.

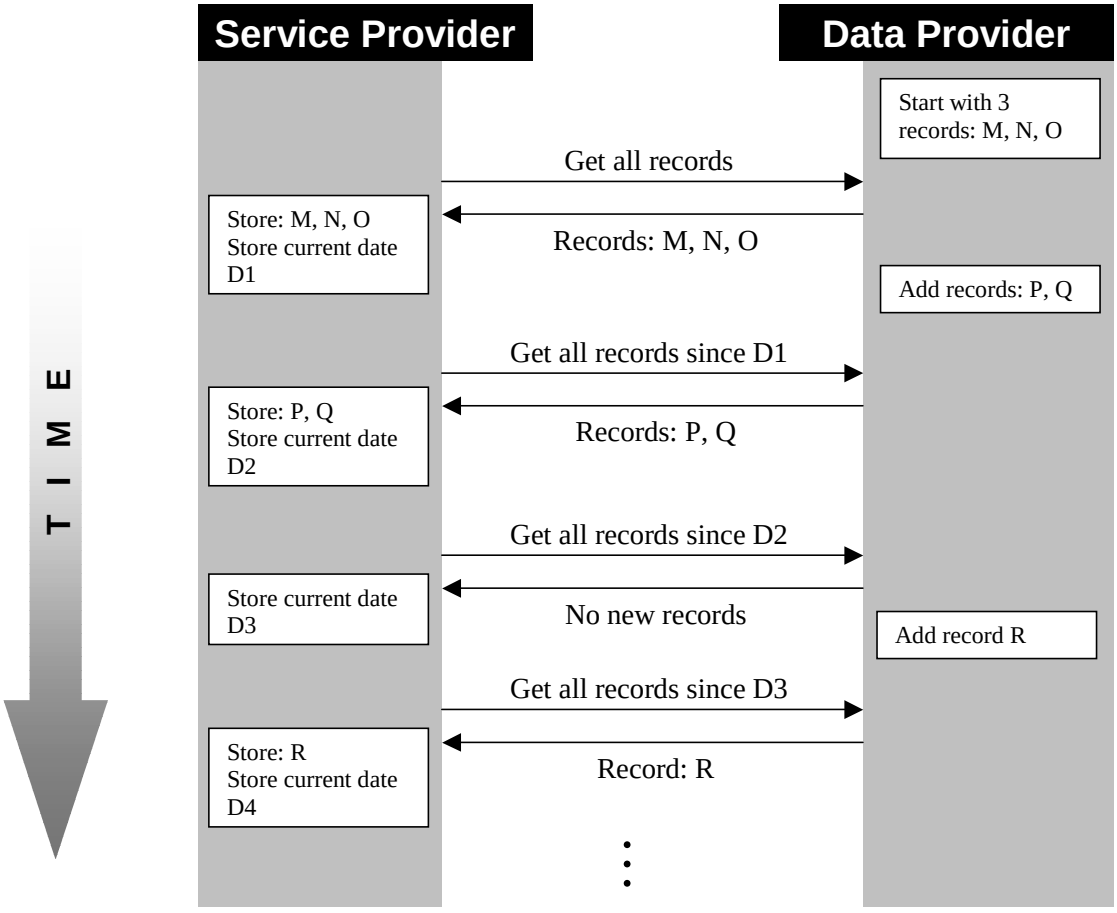

Figure 6. Example sequence of requests and responses between service and data providers

4.3 Tools and Support

The OAI website contains links to a number of useful resources that may assist developers in making their archives compliant with the protocol. The Repository Explorer is a tool that allows a user to interactively browse through an archive using only the OAI interface, while checking the interface thoroughly for errors in encoding or protocol semantics. There also is information on joining a mailing list of developers, who are more than willing to share their code and

expertise in various programming languages and on various platforms to ease the process of becoming an Open Archive. It is anticipated that a library of tools will be assembled in the near future to support new adopters of the technology.

5 OAI Support for Typical Services

5.1 Cross-Archive Searching

The most obvious service to provide would be cross-archive searching. The service provider can harvest metadata in one or more formats from multiple remote OAs and index the data according to collection, set, or specific fields within the metadata. Such an experimental search engine has already been developed at Old Dominion University (Liu, 2001) in parallel with the development of the OAI protocol.

5.2 Reference Linking

The ability to navigate quickly from one electronic publication to another that it references is a goal of many reference-linking techniques such as SFX, developed at the University Ghent (van de Sompel and Hochstenbach, 1999). OAI-accessible bibliographic metadata will greatly improve the quality and quantity of data available for constructing cross-reference databases. References could even be augmented or replaced by OAI identifiers, with an appropriate name resolution service to redirect the user to the DL that contains the referenced object.

5.3 Annotations

Since annotations are additions to existing documents, adding such a service to an existing DL usually requires the construction of a separate annotation database. In leveraging the OAI protocol, such a separate database could itself be an OA – then any entry in the OA of annotations would refer back to records in other existing OAs. A service provider would then

retrieve data from both the source OA and annotation OA before displaying the metadata to the user.

5.4 Filtering

In a profile-based filtering system, users would indicate a set of interests and then all objects corresponding to those interests would be presented to them on a continuous basis. This mode of operation is perfectly suited to the OAI protocol because of the inherently incremental nature of harvesting. Thus, a filtering or routing system could use the OAI protocol to harvest new metadata and then route that as appropriate based on a set of stored profiles.

5.5 Browsing

Unlike searching, a browsing service often requires that the metadata contain fields with controlled vocabularies that can be used to build categories within which the objects may be placed. The support for arbitrary metadata formats in the OAI protocol allows embedding of categorical data into an appropriate metadata format. In addition, the requirement for strict conformance to an XML schema can ensure that a controlled vocabulary is adhered to.

6 Existing Library Policies from an OAI Perspective

6.1 Ownership and dissemination control over digital objects and metadata

One of the major concerns that librarians have about this technology is its impact on ownership of digital objects and metadata. Some archives will openly share both with all and sundry while many archives will only share their metadata. There also are many archives that will share metadata for the purposes of building cross-archival search services but insist on users switching over to their website for the purpose of enforcing “brand recognition” or to request payment for resources. All of these scenarios are feasible since the OAI requires only that the metadata point

to the object, and this could easily be in the form of an indirect link through the originating archive. In the case of an archive that needs to restrict access to only a specified set of service providers, that can be accomplished through the access control mechanisms built into the HTTP protocol.

6.2 Changes and withdrawal

Besides ownership, most archives also reserve the right to make changes to the metadata that is associated with their digital objects. In order to propagate changes, all an archive needs to do is update the date stamp on the record so that future requests for incremental changes will result in the changed record being disseminated once again to the service provider. Service providers are expected to understand that a record received with the same identifier as a previous one is an updated version. Deletions are handled in a similar way – if identifiers for deleted records are stored at the archive, these can be returned to service providers with a special attribute that is set to indicate the record has been deleted at the source.

6.3 Preservation

Preservation of digital objects is a basic requirement of the OAI. Any archive subscribing to the OAI model of interoperability must maintain a stable collection of digital objects. The HTTP protocol has a feature to redirect URLs automatically - since objects are usually referred to by URLs, this HTTP feature can be exploited to preserve the integrity of metadata. Also, if it is expected that objects will change location often during their lifetime, they could be allocated persistent URLs (PURLs) or Handles instead of regular URLs. An essential aspect of any DL is the migration of content to newer archival technology – this is vital for interoperability efforts like the OAI since inaccessible content at a data provider will adversely affect every harvester of that data provider.

6.4 Uniqueness of objects and collections

The OAI does not require that every implementer of the harvesting protocol have a unique archive identifier. However, this is recommended so as to create a globally unique namespace for OAI identifiers. This will allow for the creation of services that are analogous to DNS name resolution – given an OAI identifier, the resolver with full knowledge of all OAs could direct a user to the archive that contains the resource.

Within archives each record must have a unique identifier so that any single GetRecord request for metadata associated with the identifier will be unambiguous.

7 Building OAI sub-Communities

7.1 Metadata formats

Communities of archives with similar interests may benefit greatly from developing their own metadata formats or simply specifying their existing metadata formats in a form that is usable with the OAI protocol. The protocol was designed to support a much higher level of semantic interoperability than is allowed by unqualified Dublin Core, so it is expected that individual archives will choose the most appropriate format for exporting their data. For example, libraries will probably use a form of MARC encoded in XML while repositories of educational resources may wish to use IMS (IMS, 1999) instead. Thus, providers of services will be able to supply users with more information, and archives will truly be able to interoperate if they have the same underlying metadata formats.

Some representatives of pre-print archives have already begun discussion of a metadata format suited for their purposes and it is hoped that this process will be initiated within other DL communities as well.

7.2 Protocol extensions

While the protocol as specified is useful for some purposes, there is no reason why an individual community cannot enhance or change the protocol to support additional features. These could take the form of either changes or additions and could be internal, with an external interface that conforms to the base protocol. Nobody expects that this protocol is a perfect solution to the problem – rather it is a stable and tested protocol that will be used for experimentation in research and production environments, leading to further evaluation and possibly newer versions after a sufficiently long period of time, set to be at least one year by the OAI. The encoding of a protocol version into the protocol further ensures that any future updates will not confuse service providers.

7.3 Shared semantics

Along with shared metadata formats a community must share a common understanding of the semantics of the metadata format. Thus, for example, if a community decides to use the RFC1807 (Lasher and Cohen, 1995) metadata format, some loosely defined fields could be further restricted for the purposes of the community, thus allowing for a more tightly coupled interoperable environment. Of course, the parallel DC metadata set must still be supported so this creates the situation where an archive may export its data in a well-defined community-specific format or a loosely defined general format satisfying the general OAI community.

7.4 Case study: Development of OAI MARC format

In a cooperative effort between Virginia Tech's Digital Library Research Laboratory and Herbert van de Sompel at Cornell University, an XML version of the US-MARC metadata format has been specified. This mapping does not attempt to encode each MARC field into a separate XML

tag, but rather encodes the fields as name/value pairs, with subfields used as required. See Figure 7 for a fragment of oai_marc XML.

```
<oai_marc xmlns="http://www.openarchives.org/OIA/oai_marc" status="n" type="a"
level="m" catForm="a">
  <fixfield id="1">"tmp96303807"</fixfield>
  <fixfield id="3">"OCoLC"</fixfield>
  <fixfield id="5">"19970728102440.0"</fixfield>
  <fixfield id="8">"971114s1996 dcu f000 0 eng d"</fixfield>
  <varfield id="35" i1="" i2="">
    <subfield label="a">1258-02760</subfield>
  </varfield>
  <varfield id="40" i1="" i2="">
    <subfield label="d">GPO</subfield>
    <subfield label="d">DLC</subfield>
    <subfield label="d">MvI</subfield>
  </varfield>
  <varfield id="49" i1="" i2="">
    <subfield label="a">VPII</subfield>
  </varfield>
  <varfield id="74" i1="" i2="">
    <subfield label="a">0378-H-12</subfield>
  </varfield>
  .
  .
  .
```

Figure 7. Fragment of sample record of XML encoding of MARC

The biggest challenges were in encoding of the character sets. Since the XML style recommended by the OAI is to use Unicode entities, all ANSEL characters need to be translated into Unicode before being exported. Composite characters also need to be changed since they are encoded differently in MARC and XML. Nevertheless, this MARC encoding in XML has generated much interest from librarians because of its simplicity and the fact that any problems can be fixed at a level outside of the schema description.

7.5 Case study: NDLTD

NDLTD, the Networked Digital Library of Theses and Dissertations, (Fox, 1999; Fox, 2001) is an international alliance of universities where students submit electronic versions of their theses and dissertations. As a preliminary step towards creating a universal catalogue of publications, the community is defining a metadata set to meet its particular needs. This metadata set is an

extension of Dublin Core with one additional field for the provision of information about the type of thesis or dissertation. The fields inherited from Dublin Core are given specific semantics that will be understood by all members of the community. Also, RDF is being investigated as an encoding strategy to incorporate links and explanations of semantics into the metadata. Ultimately, this metadata format will be exported from all NDLTD sites that are accessible through the OAI Metadata Harvesting Protocol.

8 Usage Scenarios

8.1 Dissemination of cataloguing information – MetaLibraries

In a library environment, cataloguing information is a vital resource that is shared among libraries. The OAI protocol provides a low barrier method of exchanging such cataloguing information without having to invest in high-end technology solutions. The existence of the oai_marc encoding further simplifies the task since there is now a standard way of transferring MARC records in XML.

While this may not appear very useful to large research and even public libraries, it can be very useful for smaller organizations that operate libraries. It provides a means for these smaller libraries to share their metadata with larger and peer institutions. Conceptually, it should even be possible for an appropriate organization to make available a “metalibrary” catalogue that describes every book in every OAI accessible library.

8.2 Name authority systems

The authoritativeness of names is always a problem when dealing with large quantities of data that contain references to individuals. One solution is to maintain a central (or distributed) database of names (personal and institutional) and then use links to this in each metadata item.

NDLTD has adopted this approach and is currently working with OCLC (OCLC, 2001) to set up such a system. While name information is not usually considered to be metadata, the OAI protocol can be used for name lookups by issuing GetRecord requests with the name identifier as the parameter. This is being pursued actively and illustrates a scenario where the OAI protocol can be used for simple metadata access by identifier.

8.3 Case study: NDLTD - search and classification for ETDs

NDLTD comprises a number of research universities with collections of electronic theses and dissertations. These collections are, however, managed as independent projects, very loosely linked. As an initial attempt to develop a cross-archive search service, Powell and Fox (Powell and Fox, 1998) created a federated search system. This suffered from the problem of scalability since each new archive could introduce new search semantics that would need to be integrated into the rest of the system. Also, there was no easy means of integrating the results from different systems into a single result list.

As an alternative approach, Virginia Tech is working with VTLS (VTLS, 2001) to develop a cross-archive search system based on their Virtua software. This project will use the OAI protocol to transfer metadata from individual ETD repositories into a central NDLTD collection that will be fed into Virtua and Virginia Tech's research system, MARIAN (France, 2001). In this instance, OAI technology is bridging the gaps among various different archives to increase the visibility of scholarly publications.

9 Conclusion

The Open Archives Initiative has provided the community of electronic libraries with a simple but extensible protocol to facilitate interoperability. But why do we need interoperability? The

short answer is that there are very few digital libraries that have both extensive collections and effective services. Some contain lots of data. Other provide lots of services. In either case, users do not easily find the resources related to their particular information need. Through OAI we can turn these problems into advantages by helping both data providers and service providers do a better job at their specialties, while streamlining the data provider to service provider connection. By building interoperable DLs, we can provide users with the best of both worlds, making searching of DLs a feasible notion without compromising on the quality of information management that sets digital libraries apart from the mass of data on the WWW.

10 Bibliography

ANSI/NISO. 1995. Information Retrieval (Z39.50): *Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995)*. Bethesda, MD: NISO Press.

arXiv.org. 2001. *arXiv.org e-Print archive*. <http://www.arXiv.org>.

Atkins, Anthony, Thorsten Bahne, Nune Freire, and Sarantos Kapidakis. 2001. *Interoperability Metadata Standard for Electronic Theses and Dissertations (draft)*. Available http://www.ndltd.org/standards/metadata/

Bowman, C. M., P. B. Danzig, D. R. Hardy, U. Manber, and M. F. Schwartz. 1995. The Harvest Information Discovery and Access System. *Computer Networks and ISDN Systems* 28, 119-125.

Bray, T., J. Paoli, C. M. Sperberg-McQueen, and Eve Maler, eds. 2000. *Extensible Markup Language (XML) 1.0 (Second Edition)*. W3C. Available http://www.w3.org/TR/2000/REC-xml-20001006.

CSTC. 2001. *Computer Science Teaching Center Website*. <http://www.cstc.org>.

Davis, James R., and Carl Lagoze. 2000. NCSTRL: Design and Deployment of a Globally Distributed Digital Library. *JASIS*, 51(3), 273-280.

Dublin Core Metadata Initiative. 1997. *Dublin Core Metadata Element Set Version 1.1: Reference Description*. Available http://www.dublincore.org/documents/dces/.

Fallside, David C., ed. 2000. *XML Schema*. W3C. Available http://www.w3.org/XML/Schema.

Fielding, R., J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. 1999. *Hypertext Transfer Protocol – HTTP 1.1 (RFC 2616).* Available ftp://ftp.isi.edu/in-notes/rfc2616.txt.

Fox, Edward A. 1999. Networked Digital Library of Theses and Dissertations. *Proceedings of DLW15*, July 1999. Nara, Japan: ULIS. Available http://www.ndltd.org/pubs/dlw15.doc.

Fox, Edward A. 2001. *Networked Digital Library of Theses and Dissertations*. <http://www.ndltd.org>.

France, Robert K. 2001. *MARIAN Digital Library Information System*. <http://www.dlib.vt.edu/products/marian.html>

Gravano, L., K. Chang, H. Garcia-Molina, C. Lagoze, and A. Paepcke. 1997. *STARTS: Stanford Protocol Proposal for Internet Retrieval and Search*. Available http://www-db.stanford.edu/~gravano/starts.html.

IMS Global Learning Consortium, Inc. 1999. *IMS Learning Resource Meta-data Information Model*. Available http://www.imsproject.org/metadata/mdinfov1p1.html

Kahn, Robert and Robert Wilensky. 1995. *A Framework for Distributed Digital Object Services*. Available http://www.cnri.reston.va.us/k-w.html.

Lagoze, Carl and Herbert van de Sompel. 2001. *The Open Archives Initiative Protocol for Metadata Harvesting*. Available http://www.openarchives.org/OAI/openarchivesprotocol.htm.

Lasher, R., and D. Cohen. 1995. *A Format for Bibliographic Records (RFC1807)*. Available http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1807.txt.

Liu, Xiaoming. 2001. *ARC: Cross Archive Searching Service*. <http://arc.cs.odu.edu>

OCLC, Inc. 2001. *Online Computer Library Center Website*. <http://www.oclc.org>

Open Archives Initiative. 2001. *Open Archives Initiative Website*. <http://www.openarchives.org>

Powell, James and Edward A. Fox. 1998. Multilingual Federated Searching Across Heterogeneous Collections. *D-Lib Magazine*, 4(8). Available http://www.dlib.org/dlib/september98/powell/09powell.html

Suleman, Hussein. 2001. *OAI Repository Explorer*. <http://purl.org/net/oai_explorer>

Van de Sompel, Herbert, Thomas Krichel, Michael L. Nelson and others. 2000. The UPS Prototype: An Experimental End-User Service across E-Print Archives. *D-Lib Magazine*, 6(2). Available http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html

Van de Sompel, Herbert and Carl Lagoze. 2000. The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, 6(2). Available http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

Van de Sompel, Herbert and Patrick Hochstenbach. 1999. Reference Linking in a Hybrid Libary Environment. *D-Lib Magazine*, 5(4). Available http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html

VTLS. 2001. *VTLS Website*. <http://www.vtls.com>