The Open Archives Initiative:
Building a low-barrier interoperability framework

Carl Lagoze
Digital Library Research Group
Cornell University
Ithaca, NY
+1-607-255-6046

lagoze@cs.cornell.edu

Herbert Van de Sompel
Digital Library Research Group
Cornell University
Ithaca, NY
+1-607-255-3085

herbertv@cs.cornell.edu

ABSTRACT

The Open Archives Initiative (OAI) develops and promotes interoperability solutions that aim to facilitate the efficient dissemination of content. The roots of the OAI lie in the E-Print community. Over the last year its focus has been extended to include all content providers. This paper describes the recent history of the OAI – its origins in promoting E-Prints, the broadening of its focus, the details of its technical standard for metadata harvesting, the applications of this standard, and future plans.

Categories and Subject Descriptors

D.2.12 [**Software Engineering**]: Interoperability – *Interface definition languages.*

General Terms

Experimentation, Standardization.

Keywords

Metadata, Interoperability, Digital Libraries, Protocols.

1. INTRODUCTION

In October 1999, a meeting was held in Santa Fe to discuss mechanisms to encourage the development of E-Print solutions. The group at this meeting was united in the belief that the ubiquitous interconnectivity of the Web provides new opportunities for the timely dissemination of scholarly information. The well-known physics archive run by Paul Ginsparg at Los Alamos National Laboratory has already radically changed the publishing paradigm in its respective field. Similar efforts planned, or already underway, promise to extend these striking changes to other domains.

The result of this meeting was the formation of the Open Archives Initiative (OAI) and beginning of work on a framework facilitating the federation of content providers on the Web. Since that first meeting, the OAI has undergone a period of intensive development both organizationally and technically. The original focus on E-Prints has broadened to encompass content providers from many domains (with an emphasis on what could be classified "scholarly" publishing), a refined and extensively tested technical framework has been developed, and an organizational structure to support the Initiative has been established.

The name *Open Archives Initiative* reflects the origins of the OAI in the E-Prints community where the term *archive* is generally accepted as a synonym for a repository of scholarly papers. Members of the archiving profession have justifiably noted the strict definition of an "archive" within their domain; with implications for preservation of long-term value, statutory authorization and institutional policy. The OAI uses the term "archive" in a broader sense: as a repository for stored information. Language and terms are never unambiguous and uncontroversial and the OAI respectfully requests the indulgence of the archiving community with this less constrained use of "archive".

Some explanation of the use of the term "Open" in OAI is also due. Our intention is "open" from the architectural perspective – defining and promoting machine interfaces that facilitate the availability of content from a variety of providers. Openness does not mean "free" or "unlimited" access to the information repositories that conform to the OAI technical framework. Such terms are often used too casually and ignore the fact that monetary cost is not the only type of restriction on use of information – any advocate of "free" information recognize that it is eminently reasonable to restrict denial of service attacks or defamatory misuse of information.

This paper documents the development of the Open Archives Initiative and describes the plans for the OAI for the near future. At the time of completion of this paper (May 2001), the OAI has released the technical specifications of its metadata harvesting protocol. The substantial interest in the OAI heretofore indicates that the approach advocated by the OAI – establishing a low-entry and well-defined interoperability framework applicable across domains – may be the appropriate catalyst for the federation of a broad cross-section of content providers. The coming year will indicate whether this is true and whether the technical framework defined by the metadata harvesting protocol

is a sufficient underpinning for the development of usable digital library services.

2. E-PRINT ORIGINS

The initial meeting and developments of the Open Archives Initiative are described in detail in an earlier paper [1]. This section summarizes that material from the perspective of current developments and events.

The origins of the OAI lie in increasing interest in alternatives to the traditional scholarly publishing paradigm. While there may be disagreements about the nature of what changes need to take place, there is widespread consensus that change, perhaps radical change, is inevitable. There are numerous motivating factors for this change. An increasing number of scholarly disciplines, especially those in the so-called "hard sciences" (e.g., physics, computer science, life sciences), are producing results at an increasingly rapid pace. This velocity of change demands mechanisms for reporting results with lower latency times than the ones experienced in the established journal system. The ubiquity of high-speed networks and personal computing has created further consumer demand for use of the Web for delivery of research results. Finally, the economic model of scholarly publishing has been severely strained by rapidly rising subscription prices and relatively stagnant research library budgets.

In some scholarly fields, the development of alternative models for the communication of scholarly results – many in the form of on-line repositories of EPrints – has demonstrated a viable alternative to traditional journal publication. Perhaps the best known of these is the Physics archive[1] run by Paul Ginsparg [2] at Los Alamos National Laboratory. There are, however, a number of other established efforts (CogPrints[2], NCSTRL[3], RePEC[4]), which collectively demonstrate the growing interest of scholars in using the Internet and the Web as vehicles for immediate dissemination of research findings. Stevan Harnad, among the most outspoken advocate of change, views such solutions as the first step in radical transformation of scholarly publishing whereby authors reclaim control over their intellectual property and the publishing process [3].

The October 1999 meeting in Santa Fe[5] of what was then called the UPS (Universal Preprint Service) was organized on the belief that the interoperability among these E-Print archives was key to increasing their impact. Interoperability would make it possible to bridge across, or federate, a number of archives. Issues related to interoperability are well described elsewhere [4]. It is

sufficient to say here that establishing such a framework requires both technical and organizational agreements.

There are many benefits of federation of E-Print repositories. Scholarly endeavors are increasingly multi-disciplinary and scholars should be able to move fluidly among the research results from various disciplines. Federation and interoperability also encourage the construction of innovative services. Such services might use information from various repositories and process that information to link citations, create cross-repository query interfaces, or maintain current-awareness services. The benefits of federation were demonstrated by earlier work joining the Los Alamos archive with the NCSTRL system [5], as well as in the UPS prototype [6] that was prepared for the Santa Fe meeting.

Interoperability has numerous facets including uniform naming, metadata formats, document models, and access protocols. The participants at the Santa Fe meeting decided that a low-barrier solution was critical towards widespread adoption among E-Print providers. The meeting therefore adopted an interoperability solution known as *metadata harvesting*. This solution allows E-Print (content) providers to expose their metadata via an open interface, with the intent that this metadata be used as the basis for value-added service development. More details on metadata harvesting and the OAI technical agreements are provided in Section. 4.

The result of the meeting was a set of technical and organizational agreements known as the *Santa Fe Convention.* The technical aspects included the agreement on a protocol for metadata harvesting based on the broader Dienst protocol [7], a common metadata standard for E-Prints (the Open Archives Metadata Set), and a uniform identifier scheme. The organizational agreements coming out of the meeting were informal and involved the establishment of email lists for communication amongst participants, a rudimentary registration procedure, and the definition of an acceptable use policy for consumers of harvested metadata.

The Santa Fe meeting closed with enthusiasm among the participants to refine the agreements and pursue implementation and experimentation. Within a relatively short period the technical specifications were completed and were posted on a publicly accessible web site[6] along with other results of the Santa Fe meeting. A number of the participants quickly implemented the technical agreements and others experimented with a number of prototype services.

3. BEYOND E-PRINTS

Soon after the dissemination of the Santa Fe Convention in February 2000 it became clear that there was interest beyond the E-Print community. A number of other communities were intrigued by a low-barrier interoperability solution and viewed metadata harvesting as a means to this end.

In particular, strong interest came from the research library community in the US. Key members of this community met at the so-called *Cambridge Meetings,* sponsored by the Digital Library Federation and the Andrew W. Mellon Foundation, at

---

[1] http://www.arxiv.org.

[2] http://cogprints.soton.ac.uk.

[3] http://www.ncstrl.org.

[4] http://netec.mcc.ac.uk/RePEc/.

[5] The Santa Fe meeting was sponsored by the Council on Library and Information Resources (CLIR), the Digital Library Federation (DLF), the Scholarly Publishing & Academic Resources Coalition (SPARC), the Association of Research Libraries (ARL) and the Los Alamos National Laboratory (LANL).

---

[6] http://www.openarchives.org.

Harvard University in the first half of 2000. The goal of the meetings was to explore the ways that research libraries could expose aspects of their collections to Web search engines. The participants, who included not only representatives from research libraries but also from the museum community, agreed that exposing metadata in a uniform fashion was a key step towards achieving their goal [8].

Additional evidence of the broad-based interest in Santa Fe Convention came in the form of well-attended Open Archives Initiative workshops held at the ACM Digital Library 2000 Conference in San Antonio [9] and the European Digital Library Conference in Lisbon [10]. Participants at both of these workshops included publishers, librarians, metadata and digital library experts, and scholars interested in E-Prints.

Responding to this wider interest required a reconsideration of a number of decisions made by members of the Open Archives Initiative at the Santa Fe meeting and in the months following.

- The original mission of the OAI was focused on E-Print solutions and interoperability as a means of achieving their global acceptance. While this goal was still shared by the majority of participants, it was deemed too restrictive and possibly alienating to communities not directly engaged or interested in E-Prints.

- A number of aspects of the technical specifications were specific to the original E-Print focused mission and needed to be generalized for applicability to a broad range of communities.

- The credibility of the effort was uncertain due to the lack of organizational infrastructure. Communities such as the research library community are hesitant to adopt so-called "standards" when the stability of the organization responsible for promotion and maintenance of the standard are questionable.

The issue of organizational stability was addressed first. In August 2000, the DLF (Digital Library Federation) and CNI (Coalition of Networked Information) announced organizational support and resources for the ongoing OAI effort. This support announcement was made in a press release[7] that also contained the formation of an OAI steering committee with membership from a cross-section of communities and a level of international participation (membership of the Steering Committee is listed in the Appendix in Section 7).

The OAI steering committee immediately addressed the task of compiling a new mission statement that reflected the broader scope. This mission statement is as follows:

The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. Continued support of this work remains a cornerstone of the Open Archives program. The fundamental technological framework and standards that are developing to support this work are, however, independent of the both the type of content offered and the economic mechanisms surrounding

---

[7] http://www.openarchives.org/OAISC/oaiscpress000825.htm.

that content, and promise to have much broader relevance in opening up access to a range of digital materials. As a result, the Open Archives Initiative is currently an organization and an effort explicitly in transition, and is committed to exploring and enabling this new and broader range of applications. As we gain greater knowledge of the scope of applicability of the underlying technology and standards being developed, and begin to understand the structure and culture of the various adopter communities, we expect that we will have to make continued evolutionary changes to both the mission and organization of the Open Archives Initiative.
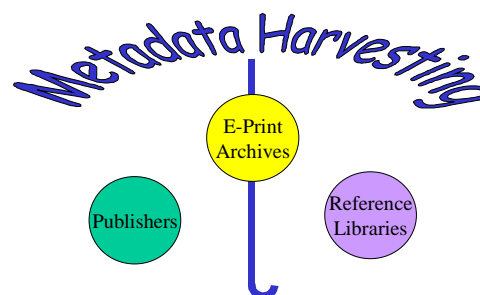
Figure 1 - A framework for multiple communities

This mission is illustrated in Figure 1, where the technical framework is designed as an umbrella that can be exploited by a variety of communities.

A key element of this mission statement is the formulation of the OAI as an experiment: "*an organization and an effort explicitly in transition*". The organization is well aware that the technical infrastructure it proposes – metadata harvesting – has to be proven as an effective means of facilitating interoperability or even what that interoperability will achieve. The organizational structure and strategy reflects a belief among the steering committee that "proving the concept" will require a delicate balance between stability and flexibility. Furthermore, there is strong consensus that goals and scope of the OAI should be controlled – while interoperability is a wide-open area with many potential areas of investigation, the OAI should resist expanding its scope until its current technical goals are met and justified.

The Steering Committee also took steps to address the E-Print focus of the Santa Fe technical agreements and to fix other problems that were revealed in testing of those agreements. A technical committee was formed and a meeting organized at Cornell University in September 2000. The results of that meeting are reported in the next section.

4. TECHNICAL FRAMEWORK

The technical framework of the Open Archives Initiative is intended to provide a low-barrier approach to interoperability. The membership of the OAI recognizes that there are functional limitations to such a low-barrier framework and that other interoperability standards, for example Z39.50, address a number of issues in a more complete manner. However, as noted by Bill

Arms [11], interoperability strategies generally increase in cost (difficulty of implementation) with an increase in functionality. The OAI technical framework is not intended to replace other approaches but to provide an easy-to-implement and easy-to-deploy alternative for different constituencies or different purposes than those addressed by existing interoperability solutions. As noted earlier, experimentation will prove whether such low-barrier interoperability is realistic and functional.

At the root of the technical agreement lies a distinction between two classes of participants:

- *Data Providers* adopt the OAI technical framework as a means of exposing metadata about their content.

- *Service Providers* harvest metadata from data providers using the OAI protocol and use the metadata as the basis for value-added services.

The remainder of this section describes the components of this technical framework. A theme carried through the framework and the section is the attempt to define a common denominator for interoperability among multiple communities while providing enough hooks for individual communities to address individual needs (without interfering with interoperability). More details on the technical framework are available in the Open Archives Metadata Protocol specification available at the OAI web site.[8]

4.1 Metadata

The OAI technical framework addresses two well-known metadata requirements: interoperability and extensibility (or community specificity). These issues have been a subject of considerable discussion in the metadata community [12, 13] – the OAI attempts to answer this in a simple and deployable manner.

The requirement for metadata interoperability is addressed by requiring that all OAI data providers supply metadata in a common format – the Dublin Core Metadata Element Set [14]. The decision to mandate a common element set and to use the Dublin Core was the subject of considerable discussion in the OAI. One approach, which has been investigated in the research literature [15], is to place the burden on the consumer of metadata rather than the provider, tolerating export of heterogeneous metadata and relying on services to map amongst the representations. OAI, however, is purposely outside the domain of strict research, and in the interest of easy deployment and usability it was decided that a common metadata format was the prudent decision.

The decision to use the Dublin Core was also the result of some deliberation. The original Santa Fe convention took a different course – defining a metadata set, the Open Archives Metadata Set, with some functionality tailored for the E-Print community. The broadening of the focus of the OAI, however, forced reconsideration of this decision and the alternative of leveraging the well-known and active work of the DC community in formulating a cross-domain set for resource discovery was chosen.

Those familiar with the Dublin Core will note that all fields in DC are *optional*. The OAI discussed requiring a number of DC elements in OAI records. While such requirement might be

---

[8] http://www.openarchives.org/OAI/openarchivesprotocol.htm.

preferable from the perspective of interoperability, the spirit of experimentation in the OAI persuaded the committee to keep all elements optional. The committee agreed that it would be desirable at this early stage to encourage metadata suppliers to expose DC metadata according to their own needs and thereby reveal a market of community-developed metadata practices.

It should be noted that the specific decision was to use *unqualified* Dublin Core as the common metadata set. This decision was made based on the belief that the common metadata set in OAI is explicitly purposed for coarse granularity resource discovery. As discussed elsewhere [16], qualification of DC, for the purpose of more detailed *description* (rather than simple *discovery*) is still an area of some contention and threatens to interfere with the goal of simple resource discovery. The OAI takes the approach of strictly separating simple discovery from community-specific description.

Community-specific description, or metadata specificity, is addressed in the technical framework by support for parallel metadata sets. The technical framework places no limitations on the nature of such parallel sets, other than that the metadata records be structured as XML documents, which have a corresponding XML schema for validation (as described in section 4.2). At the time of completion of this paper (January 2001), initial steps have been taken to encourage the development of community-specific harvestable metadata sets. Representatives of the E-Print community have been working on a metadata set targeted at the E-Print community under the name EPMS. Representatives of the research library community have proposed a similar effort and there are calls for proposals from other communities (e.g., the museum community, Open Language Archives).

4.2 Records, Repositories, and Identifiers

The OAI technical framework defines a *record*, which is an XML-encoded byte stream that serves as a packaging mechanism for harvested metadata. A record has three parts:

- *header* – containing information that is common to all records (it is independent of the metadata format disseminated in the record) and that is necessary for the harvesting process. The information defined in the header is the unique identifier for the record (described below), and a datestamp indicating the date of creation, deletion, or latest date of change in the metadata in the record.

- *metadata* – containing metadata in a single format. As noted in section 4.1, all OAI data providers must be capable of emitting records containing unqualified DC metadata. Other metadata formats are optional.

- *about* – an optional container to hold data about the metadata part of the record. Typically, the "about" container could be used to hold rights information about the metadata, terms and conditions for usage of the metadata, etc. The internal structure of the "about" container is not defined by the protocol. It is left to individual communities to decide on its syntax and semantics through the definition of a schema.

A sample OAI record is shown in Figure 2.

Metadata records are disseminated from *Repositories*, which are network accessible servers of data providers. An OAI-

conformant repository supports the set of OAI protocol requests defined in Section 4.4. Abstractly, repositories contain *items*, and each metadata record harvested from a repository corresponds to an item. (There is a many-to-one relationship of records to items, since metadata can be expressed in multiple formats). The nature of an item - for example, what type of metadata is actually stored in the item, what type is derived on the fly, and whether the item includes the "full content" described by the metadata - is outside the scope of the OAI protocol. This admittedly indistinct nature of an item is intentional. The OAI harvesting protocol is meant to be agnostic as to the nature of a data provider – it supports those that have content with fixed metadata records, those that computationally derive metadata in various formats from some intermediate form or from the content itself, or those that are metadata stores or

```
<header>
    <identifier>oai:arXiv:9901001</identifier>
    <datestamp>1999-01-01</datestamp>
</header>
<metadata>
  <dc xmlns="http://www.openarchives.org/OAI/dc.xsd">
    <title>Quantum slow motion</title>
    <creator>Hug, M.</creator>
    <creator>Milburn, G. J.</creator>
    <date>1999-01-01</date>
    <type>e-print</type>
    <identifier>http://arXiv.org/abs/9901001</identifier>
  </dc>
</metadata>
<about>
  <dc xmlns=" httpd://www.openarchives.org/OAI/dc.xsd">
     <rights>Metadata may be used without restrictions</rights>
  </dc>
</about>
```

Figure 2 – Sample OAI Record

metadata intermediaries for external content providers.

As illustrated in Figure 2 each record has an identifier. The nature of this identifier deserves some discussion. Concretely, the record identifier serves as a key for extracting metadata from an item in a repository. This key, parameterized by a metadata format identifier, produces an OAI record. Since the identifier acts in this manner as a key it must be unique within the repository; each key corresponds to metadata derived from one item. The protocol itself does not address the issues of inter-repository naming or globally unique identifiers. Such issues are addressed at the level of registration, which is described in Section 4.5

The record identifier is expressly *not* the identifier of the item – the issue of identifiers for contents of repositories is intentionally outside the scope of the OAI protocol. Undoubtedly, many clients of the OAI protocol will want access to the full content

described by a metadata record. The protocol recommends that repositories use an element in metadata records to establish a linkage between the record and the identifier (URL, URN, DOI, etc.) of the associated item. The mandatory Dublin Core format provides the *identifier* element that can be used for this purpose.

4.3 Selective Harvesting

A protocol that only enabled consumers of metadata to gather all metadata from a data provider would be cumbersome. Imagine the transactions with large research libraries that expose the metadata in their entire catalog through such a protocol!

Thus, some provision for selective harvesting, which makes it possible in the protocol to specify a subset of records to be harvested, is desirable. Selection, however, has a broad range of functionality. More expressive protocols include provisions for the specification of reasonably complete predicates (in the manner of database requests) on the information requested. The OAI decided that such high functionality was not appropriate for a low-barrier protocol and instead opted for two relatively simple criteria for selective harvesting.

- *Date-based* – As noted in Section 4.2, every record contains a date stamp, defined as "the date of creation, deletion, or latest date of modification of an item, the effect of which is a change in the metadata of a record disseminated from that item". Harvesting requests may correspondingly contain a date range for harvesting, which may be total (between two dates) or partial (either only a lower bound or an upper bound). This date-based harvesting provides the means for incremental harvesting. For example, a client may have a weekly schedule for harvesting records from a repository, and use the date-based selectivity to only harvest records added or modified since the last harvesting time.

- *Set-based* – The protocol defines a *set* as "an optional construct for grouping items in a repository for the purpose of selective harvesting of records". Sets may be used in harvesting requests to specify that only records within a specific grouping should be returned from the harvesting request (note that each item in a repository may be organized in one set, several sets, or no sets at all). Each repository may define a hierarchical organization of sets that can have several top-level nodes, each of which is a *set*. Figure 3 illustrates a sample set hierarchy for a fictional E-Print repository. The actual meaning of the sets is not defined within the protocol. Instead, it is expected that communities that use the OAI protocol may formulate well-defined set configurations with perhaps a controlled vocabulary for set names, and may even develop mechanisms for exposing these to service providers. As experiments with the OAI protocol proceed in the future, it will be interesting to see how communities exploit the set mechanism and if it provides sufficient functionality.

Even with the provisions for selective harvesting, it is possible that clients will make harvesting requests of repositories that are large and burdensome to fulfill in a single response. Some other protocols make provision for such cases with the notion of *state*

Institutions
        Cornell University
        Virginia Tech
Subjects
        Computer Science
        High Energy Physics

Figure 3 - Sample Set Hierarchy

and *result sets* – a client explicitly opens a session, conducts transactions within that session, and then closes a session. Yet, session maintenance is notably complex and ill suited for protocols such as HTTP, which is intended as the carrier protocol for OAI requests and responses. Instead the OAI uses a relatively simple flow control mechanism that makes it possible to partition large transactions among several requests and responses. This flow control mechanism employs a *resumption token*, which is returned by a repository when the response to a harvesting request is larger than the repository may wish to respond to at one time. The client can then use the resumption token to make subsequent requests until the transaction is complete.

4.4 Open Archives Metadata Harvesting Protocol

The initial protocol that came out of the Santa Fe meeting was a subset of the Dienst protocol. While that subset protocol was functionally useful for metadata harvesting, aspects of its legacy context presented barriers to simple implementation. The current technical framework is built around a more focused and easier to implement protocol – the Open Archives Metadata Harvesting Protocol.

The Open Archives Metadata Harvesting Protocol consists of six requests or verbs. The protocol is carried within HTTP POST or GET methods. The intention is to make it simple for data providers to configure OAI conformant repositories by using readily available Web tools such as libwww-perl[9]. OAI requests all have the following structure:

- *base-url* – the Internet host and port of the HTTP server acting as a repository, with an optional path specified by the respective HTTP server as the handler for OAI protocol requests.

- *keyword arguments* – consisting of a list of key-value pairs. At a minimum, each OAI protocol request has one key=value pair that specifies the name of the OAI protocol request.

Figure 4 shows the encoding of a sample OAI protocol request using both HTTP GET and POST methods. The request is the *GetRecords* verb, and the specific example requests the return of the record with identifier *oai:arXiv:hepth01* in *dc* (Dublin Core) format.

```
GET Request

http://ana.oa.org/OAI-script?
        verb=GetRecord&
        identifier=oai:arXiv:hep-th01&
        metadataPrefix=dc

POST Request

POST http://an.oa.org/OAI-script
Content-Length: 62
Content-Type: application/x-www-form-urlencoded
verb=GetRecord&
identifier=oai:arXiv:hep-th01&
metadataPrefix=dc
```

Figure 4 - Sample OAI Request Encoding

The response to all OAI protocol requests is encoded in XML. Each response includes the protocol request that generated the response, facilitating machine batch processing of the responses. Furthermore, the XML for each response is defined via an XML schema. [17-19]. The goal is to make conformance to the technical specifications as machine verifiable as possible – a test program should be able to visit an OAI repository, issue each protocol request with various arguments, and test that each response conforms to the schema defined in the protocol for the response.

The remainder of this section summarizes each of the protocol requests.

4.4.1 GetRecord
This verb is used to retrieve an individual record (metadata) from an item in a repository. Required arguments specify the identifier, or key, of the requested record and the format of the metadata that should be included in the record.

4.4.2 Identify
This verb is used to retrieve information about a repository. The response schema specifies that the following information should be returned by the *Identify* verb:

- A human readable name for the repository.

- The base URL of the repository.

- The version of the OAI protocol supported by the repository.

- The e-mail address of the administrator of the repository.

In addition to this fixed information, the protocol provides a mechanism for individual communities to extend the functionality of this verb. The response may contain a list of *description* containers, for which a community may define an XML schema that specifies semantics for additional description of the repository.

4.4.3 ListIdentifier
This verb is used to retrieve the identifiers of records that can be harvested from a repository. Optional arguments permit selectivity of the identifiers - based on their membership in a specific set in the repository or based on their modification, creation, or deletion within a specific date range.

4.4.4 ListMetadataFormats
This verb is used to retrieve the metadata formats available from a repository. An optional argument restricts the request to the formats available for a specific record.

4.4.5 ListRecords
This verb is used to harvest records from a repository. Optional arguments permit selectivity of the harvesting - based on the membership of records in a specific Set in the repository or based on their modification, creation, or deletion within a specific date range.

4.4.6 ListSets
This verb is used to retrieve the set structure in a repository.

4.5 Data Provider Conformance and Registration

The OAI expects that data providers will fall into three layers of participation, each higher layer implying the preceding layer(s);

1) *OAI-conformant* – These are data providers who support the protocol definition. As stated earlier, conformance is testable since there are XML-schemas to validate all responses. No doubt, the OAI will not be able to track every provider using the protocol since use of it does not require any licensing or registration procedure.

2) *OAI-registered* – These are data providers who register in an OAI-maintained database, which will be available through the OAI web site. Registration will entail that the data provider gives a BASE-URL, which the registration software will then use to test compliance.

3) *OAI-namespace-registered* – These are data providers who choose to name their records in conformance with an OAI naming convention for identifiers. Names that follow this convention have the following three components:

   a) *oai* – A fixed string indicating that the name is in the OAI namespace.

   b) *<repoID>* - An identifier for the repository that is unique within the OAI namespace.

   c) *<localID>* - An identifier unique within the respective repository.

   An example of a name that uses this naming scheme is: *oai:arXiv:hep-th01*

The advantage for repositories of adopting this naming convention is that record identifiers will be resolvable via a central OAI resolution service, that will be made available at the OAI web site. The intention is to make this resolver OpenURL-aware[10], as a means to support open linking [20-22] based on OAI identifiers. The attractiveness of such an approach has been demonstrated in an experiment conducted in the DOI namespace[11]. A process for fast-track standardization of OpenURL has recently started with NISO.

5. TESTING AND REFINEMENT

Participants of the September 2000 technical meeting at Cornell developed a rough outline of the technical framework. However, the task of normalizing and putting the framework into the form of a specification was undertaken at Cornell University, where Open Archives Initiative activities are coordinated. Continuous feedback from the alpha-test group that implemented consecutive versions of the protocol played an important role in this activity. Participants in the alpha-test group were solicited from both the original Santa Fe Convention E-Print community and from the attendees at the DLF sponsored Cambridge meetings. These solicitations led to a quite comprehensive and diverse testing community, organized around an alpha testers email list. The complete list of alpha testers is shown in the Appendix in section

---

8. It includes representatives from the E-Print community, museums, research libraries, repositories of publisher metadata, and collectors of web site metadata. In addition, the alpha test included two rudimentary service providers who constructed search interfaces based on metadata harvested from the OAI-conformant alpha testers.

Three alpha-testers deserve special mention. Hussein Suleman at Virginia Tech created and continuously updated a repository explorer that allowed alpha-testers to examine the compliance of their repositories to the most recent version of the protocol document. Simeon Warner (Los Alamos National Laboratory and arXiv.org) and Michael Nelson (University of Northern Carolina and NASA) did extensive proofreading of new versions of the protocol document before release to the alpha-group.

The results of these tests are quite encouraging. First, the protocol specification has passed through a number of revisions and has been vetted extensively for errors and ambiguities. Second, virtually all the testers remarked at the ease of implementation and conceptual simplicity of the protocol.

6. THE ROAD AHEAD

At the beginning of 2001 the Open Archives Initiative began the next phase of its work, public deployment and experimentation with the technical framework. To initiate this process two public meetings were scheduled. A US meeting was be held in Washington DC on January 23. Registration for this meeting was closed when the maximum of 140 participants was reached. The participants represented a wide variety of communities. A European meeting was scheduled for February 26 in Berlin. Participants at these meetings heard a complete overview of the goals of the OAI and the particulars of the technical framework. The meetings also provided an opportunity for the development of communities within the Open Archives framework. Communities may take the form of groups of data providers that exploit the extensibility of the Open Archives Harvesting Protocol to expose purpose-specific metadata or the development of targeted services.

These meetings were meant to be a "kick off" for an extended period of experimentation (at least one year) with the harvesting protocol. The OAI intends during this period to keep the protocol as stable as possible. This experimentation phase is motivated by the belief that the community needs to fully understand the functionality and limits of the interoperability framework before considering major changes or expansion of functionality.

During this experimentation period, three large research and implementation projects, in the U.S. and in Europe, plan to experiment with the functionality of the OAI technical framework:

1. *National Science Digital Library*[12] – NSDL is a multi-participant project in the US funded by the National Science Foundation with the goal of creating an online network of learning environments and resources for science, mathematics, engineering, and technology education. Our group at Cornell is funded under the core infrastructure portion of the NSDL. In the context of the OAI alpha

---

[10] http://www.sfxit.com/OpenURL

[11] http://sfxserv.rug.ac.be:8888/public/xref/

[12] http://www.ehr.nsf.gov/ehr/due/programs/nsdl/.

testing we experimented with a harvesting service that will later form the basis for other services in our NSDL infrastructure. Future plans include working with our partners at the San Diego Super Computer Center to harvest metadata via OAI, post-process and normalize it for storage in a metadata repository, and then make that metadata searchable using the SDLIP [23] protocol.

2. *Cyclades*[13] - This is a project funded by the European Commission with partners in Italy, Germany, and the U.K. The main objective of Cyclades is to develop advanced Internet accessible mediator services to support scholars both individually and as members of communities when interacting with large interdisciplinary electronic archives. Cyclades plans to investigate the construction of these services on the Open Archive foundation.

3. *Digital Library Federation Testbed* – As a follow-up to the Cambridge meetings (described earlier in this paper) a meeting[14] of interested project participants was convened in October 2000 by The Andrew W. Mellon Foundation to explore technical, organizational, and resource issues for broad-based metadata harvesting and to identify possible next steps. The result of this meeting was the commitment by a number of institutions (research libraries, other information-based organizations) to expose metadata from a number of collections through the Open Archives technical infrastructure, and experiment with services that use this metadata.

These projects promise to expose what is possible using the OAI framework and how it might be changed or expanded.

We are intrigued by the period of experimentation that lies ahead and encouraged by the widespread interest in the Open Archives Initiative. In the context of this enthusiasm, we are aware of the need to be circumspect in our approach towards the OAI and its goals. As Don Waters, a member of the OAI Steering Committee, pointed out, the technical proposals of the OAI include a number of assumptions about issues not yet fully understood.

- What is the value of a common metadata set?

- What are the interactions of native metadata set with the minimal, conventional set?

- What are the incentives and rewards for institutions and organizations for participating in such a framework?

- What are the intellectual property issues vis-à-vis harvestable metadata?

- Will this technical framework encourage new models of scholarly communication?

These, and many other questions, are all in need of thorough examination. Too often members of the digital library community have made casual statements that "interoperability is good", "metadata is important", and that "scholarly publishing is changing". At the minimum, we hope that the OAI will create a framework for serious investigation of these issues and lay the foundation for more informed statements about the issues critical to the success of our field.

7. Appendix A – OAI STEERING COMMITTEE

Names are followed by affiliations:

- Caroline Arms (Library of Congress)
- Lorcan Dempsey (Joint Information Systems Committee, UK)
- Dale Flecker (Harvard University)
- Ed Fox (Virginia Tech)
- Paul Ginsparg (Los Alamos National Laboratory)
- Daniel Greenstein (DLF)
- Carl Lagoze (Cornell University)
- Clifford Lynch (CNI)
- John Ober (California Digital Library)
- Diann Rusch-Feja (Max Planck Institute for Human Development)
- Herbert Van de Sompel (Cornell University)
- Don Waters (The Andrew W. Mellon Foundation)

8. Appendix B – ALPHA TEST SITES

The following institutions participated in the alpha testing of the technical specifications:

- CIMI Museum Consortium
- Cornell University
- Ex Libris
- Los Alamos National Laboratory
- NASA
- OCLC
- Old Dominion University
- UKOLN Resource Discovery Network
- University of Illinois Urbana Champaign
- University of North Carolina
- University of Pennsylvania
- University of Southampton
- University of Tennessee
- Virginia Tech

9. ACKNOWLEDGMENTS

---

[13] http://cinzica.iei.pi.cnr.it/cyclades/,

[14] http://www.clir.org/diglib/architectures/testbed.htm.

10. REFERENCES

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

[2] Ginsparg, P., *Winners and Losers in the Global Research Village.* The Serials Librarian, 1997. **30**(3/4): p. 83-95.

[3] Harnad, S., *Free at Last: The Future of Peer-Reviewed Journals.* D-Lib Magazine, 1999. **5**(12).http://www.dlib.org/dlib/december99/12harnad.html.

[4] Paepcke, A., *et al.*, *Interoperability for Digital Libraries Worldwide*, *Communications of the ACM*. 1998, **41**(4): 33-42.

[5] Halpern, J.Y. and C. Lagoze. *The Computing Research Repository: Promoting the Rapid Dissemination and Archiving of Computer Science Research.* In *Digital Libraries '99, The Fourth ACM Conference on Digital Libraries*. 1999. Berkeley, CA.

[6] Van de Sompel, H., T. Krichel, and M.L. Nelson, *The UPS Prototype: an experimental end-user service across e-print archives*, in *D-Lib Magazine*. 2000.

[7] Lagoze, C. and J.R. Davis, *Dienst - An Architecture for Distributed Document Libraries.* Communications of the ACM, 1995. **38**(4): 47.

[8] *A New Approach to Finding Research Materials on the Web*, . 2000, Digital Library Federation. http://www.clir.org/diglib/architectures/vision.htm.

[9] Anderson, K.M., *et al.*, *ACM 2000 digital libraries : proceedings of the fifth ACM Conference on Digital Libraries, June 2-7, 2000, San Antonio, Texas.* 2000, New York: Association for Computing Machinery. xiii, 293.

[10] Borbinha, J. and T. Baker, *Research and advanced technology for digital libraries: 4th European Conference, ECDL 2000, Lisbon, Portugal, September 18-20, 2000 : proceedings.* Lecture notes in computer science 1923. 2000, Berlin ; New York: Springer. xvii, 513.

[11] Arms, W.Y., *Digital libraries.* Digital libraries and electronic publishing. 2000, Cambridge, Ma.: MIT Press.

[12] Lagoze, C., C.A. Lynch, and R.D. Jr., *The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata.* 1996, Cornell University Computer Science. http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593.

[13] Lagoze, C. *Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience.* in *Seminar on Metadata*. 2000.

Archiefschool, Netherlands Institute for Archival Education and Research, The Hague.

[14] Weibel, S., *The Dublin Core: A simple content description format for electronic resources.* NFAIS Newsletter, 1998. **40**(7): p. 117-119.

[15] Chang, C.-C.K. and H. Garcia-Molina. *Mind your vocabulary: query mapping across heterogeneous information sources.* In *International Conference on Management of Data and Symposium on Principles of Database Systems*. 1999. Philadelphia: ACM: 335-346.

[16] Lagoze, C. and T. Baker, *Keeping Dublin Core Simple.* D-Lib Magazine, 2001. **7**(1).

[17] Fallside (ed.), D.C., *XML Schema Part 0: Primer*. 2000, World Wide Web Consortium. http://www.w3.org/TR/xmlschema-0/.

[18] Thompson, H.S., *et al.*, *XML Schema Part 1: Structures*. 2000, World Wide Web Consortium. http://www.w3.org/TR/xmlschema-1/.

[19] Biron, P.V. and A. Malhotra, *XML Schema Part 2: Datatypes*. 2000, World Wide Web Consortium. http://www.w3.org/TR/xmlschema-2/.

[20] Van de Sompel, H. and P. Hochstenbach, *Reference Linking in a Hybrid Library Environment , Part 3: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment.* D-Lib Magazine, 1999. **5**(10).http://www.dlib.org/dlib/october99/van_de_sompel/10van_de_sompel.html.

[21] Van de Sompel, H. and P. Hochstenbach, *Reference Linking in a Hybrid Library Envronment:, Part 1: Frameworks for Linking.* D-Lib Magazine, 1999. **5**(4).http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html.

[22] Van de Sompel, H. and P. Hochstenbach, *Reference Linking in a Hybrid Library Environment, Part 2: SFX, a Generic Linking Solution*, in *D-Lib Magazine*. 1999.

[23] Paepcke, A., *et al.*, *Search Middleware and the Simple Digital Library Interoperability Protocol.* D-Lib Magazine, 2000. **5**(3).http://www.dlib.org/dlib/march00/paepcke/03paepcke.html.