Report on Open Archives Initiative Technical Committee Meeting

Ithaca NY, 7-8 September 2000

Context

The original aim of the Open Archives Initiative was to provide an infrastructure for interoperability among sites supporting author self-archiving and thereby promote their wide acceptance. Although the Initiative generally concentrated on technical matters, its mission reflected its roots in the e-print community and the underlying political agenda to promote the ongoing transformation of scholarly communication. The inaugural meeting of the Open Archives Initiative (OAI) in October 1999 spawned an agreement now known as the Santa Fe Convention.

The Santa Fe Convention is a set of relatively simple interoperability agreements that facilitate a minimal but potentially highly functional level of interoperability among scholarly e-print archives through metadata harvesting. The interoperability agreements are a combination of organizational principles and technical specifications The Convention gives *data providers* -- individual archives -- relatively easy-to-implement mechanisms for making metadata in their archives externally available. This external availability then makes it possible for *service providers* to build higher levels of functionality, mediator services, using the information made available from scholarly archives that adopt the convention. These services may combine and process information from individual archives and then may offer increased functionality to support discovery, presentation and analysis of data originating from compliant archives.

Since the publication of the Santa Fe Convention in February 2000, interest has emerged from other communities who are interested in applying the framework for a wide variety of scholarly materials beyond e-prints. In order to respond to this wider interest, the OAI undertook a number of actions:

The technical specifications were reconsidered in response to comments that certain aspects were e-print specific. Experimentation and discussion in the original e-print community also identified elements in the original specifications that required reconsideration.

The original e-print specific mission statement was reconsidered. Rather than focusing on a political agenda focusing on author self-archiving, OAI's mission was reformulated to supply and promote an application independent technical framework - a supportive infrastructure that empowers different scholarly communities to pursue their own interests in interoperability in the technical, legal, business, and organizational contexts that are appropriate to them.

Organizational changes were instituted to provide stability and credibility to the wider community base. A Steering Committee was appointed with the task of overseeing the pursuit of the mission. The activities of the Steering Committee will receive support from both the Digital Library Federation and the Coalition for

Networked Information. In addition a Technical Committee was formed to focus on generalization and stabilization of the technical framework.

The Cornell Meeting of the OAI Technical Committee

A meeting of the OAI Technical Committee was held on September 7-8 2000 at Cornell University to revise the Santa Fe Convention in light of the changed context. The meeting included analysis of the current and emerging use of the interoperability framework and initiated the process of upgrading it to better serve the needs of a more general user base.

The meeting set out by agreeing on the following issues:

- The OAI interoperability framework should no longer only be concerned with e-prints, but with scholarly data-archives in general.
- Most fundamental principles of the Santa Fe Convention [open, harvestable archives ; data provider & service provider model ; managed archives] can be maintained in the extended scope. One concept [the definition of a record in an archive] should be reconsidered during the meeting.
- Most abstract principles that are presented in the Santa Fe Convention [metadata harvesting ; OAI namespace ; acceptable use ; registration of data providers, service providers and metadata formats] can be maintained in the extended scope. One concept [shared metadata set & parallel metadata sets] should be reconsidered during the meeting.
- The existing technical implementation of these abstract principles should be reconsidered during the meeting because of the extension of scope and because of experiences with actual implementations.

The goal of the meeting was development of a new set of technical guidelines for consideration by the Open Archives Steering Committee and ultimate public dissemination by the beginning of 2001. Recognizing that any specification is subject to review and refinement, the attendees attempted to develop specifications that were:

- Stable for experimentation;
- Low risk for early adopters;
- Sufficiently easy implement so as to optimize the chances for future interoperability across communities.

It was decided to discontinue the name “Santa Fe Convention”. The new name for the interoperability specification is “the Open Archives Harvesting Framework Specifications”.

A formal specification document is currently being developed. The new specifications will be disseminated to the public in January 2001. Documentation, accompanying tools and software will be produced in parallel.

The remainder of this document summarizes the Open Archives Harvesting Framework Specifications. It focuses on issues where the Open Archives Harvesting Framework differs from the specifications in the Santa Fe Convention.

Record in an archive

The ambiguity in the original agreement about the definition of a *record* has been clarified. A record in an archive has been defined to be a metadata-record. The metadata record describes – and can contain an entry point to – full-content.

Metadata

The requirement to have a shared, basic metadata set to facilitate interoperability across communities was reconfirmed. Also the notion of parallel metadata sets that serve specific needs of communities and archives was reconfirmed.

The shared metadata set developed during the original Santa Fe meeting -- the OAMS -- was deemed inappropriate for cross-community use due to some e-print specific aspects. Instead, the Dublin Core Element Set was selected as the common metadata set. This selection leverages years of work in the Dublin Core Metadata Initiative in developing cross-community consensus.

Initial steps were taken to encourage the development of community-specific harvestable metadata sets. Representatives of the e-print community at the meeting decided to propose a metadata set targeted at the e-print community under the name EPMS by the beginning of 2001. Representatives of the research library community proposed a similar effort and calls for proposals from other communities (e.g., the museum community, Open Language Archives) will be issued.

To distinguish between metadata specific to harvesting functionality and other metadata (both shared and community specific), a carrier syntax in XML was developed. This syntax transports packages of specific sets (e.g., Dublin Core, EPMS) within a contextual wrapper that contains metadata specific to the harvesting interactions between data and service providers.

Identifiers and an OAI namespace

The concept of unique identifiers within an OAI namespace has been maintained. Its implementation has been revised for compliance with general URI principles and to allow for the building of resolution mechanisms and services across OAI-compliant archives. The following identifier syntax is proposed:

full-identifier = oai : archive-identifier : record-identifier

Where:

oai - the scheme (which will be registered as a URI scheme)

archive-identifier - the unique identifier of an archive (which will be registered within the Open Archives Initiative)

record-identifier - the unique, persistent identifier of a record within the archive (the syntax of this name is archive-specific within the limitations of the URI syntax).

Sets (formally called Partitions)

The Partition concept in the original technical agreement has been retained, but has been renamed Sets. The concept allows individual records in archives to be arranged in unconstrained sets at the discretion of the archive administrator. These Sets can then be organized in a hierarchical fashion to expose the internal structure of the archive. Individual communities can make explicit agreements on the actual meaning of Sets within their communities. As such, individual communities may use Sets as a tool for selective harvesting. However, they are not meant to serve as a general tool for determining categories.

OAI Harvesting Protocol

At the heart of the technical agreements of the OAI is the metadata harvesting protocol, which provides a simple interface to transfer metadata from a data provider to a service provider. The original specifications for this interface in the Santa Fe Convention were a subset of the more expressive Dienst protocol (called the Open Archives Dienst Subset). Discussions at the meeting revealed that while the semantics of the subset service requests were generally correct, many of the artifacts of the broader Dienst protocol presented unnecessary complexities to implementers. As a result an independent OAI protocol was developed, derived from the original Open Archives Dienst Subset.

This protocol contains the following service requests:

Identify – returns a self-description of the archive, containing information submitted at time of registration and other administrative information;

ListMetadataFormats – returns a list of identifiers of metadata formats that are offered by the archive in general or for a particular record;

ListSets – returns a structured list of sets (formally called partitions) within which records may be located;

ListRecords – returns a list of record identifiers, and optionally metadata, within a specified range of dates and/or a specified Set. Flow control is achieved by the use of server-generated continuation tokens;

GetRecord – returns the metadata associated with an identifier.

These service requests and their parameters are encoded into standard HTTP URIs. Responses to the requests are XML documents. This allows for simple implementation using CGI scripts or similar technology for data providers while service providers can exploit the proliferation of XML parsers to ease the harvesting of data.

Registration

It was decided that the process of registration should become more automated. Also, it was decided that the OAI should currently keep registration of compliant data providers (archives), compliant service providers and metadata formats under its own governance.

Information elements that need to be included for the registration of a data provider have been reconsidered in light of the extension of the scope of the Initiative. The existing registration via the provision of a data provider template will be replaced by:

- On-line registration of an archive identifier;
- On-line registration of the BASE-URL of the archive's OAI Protocol implementation;
- Support by the archive of the Identity verb that will expose essential information about the archive's machine interface, its policies, etc.

Registration of metadata formats – including format identifier, description of the format's semantics and the DTD of its XML transportation format -- will be automated and will include a check of the validity of the DTD.

Registration of service providers has not been discussed. A revision of the information elements required for registration will be proposed.

Acceptable Use

Acceptable use of data

The "gentlemen's agreement" between data providers and service providers, whereby

- The data providers expresses usage restrictions for data harvested from its archive;
- The service provider expresses to comply with those restrictions;

is maintained. However, an explicit distinction should be made between the harvesting of metadata -- which is the topic of the specifications -- and the harvesting of the full-content – which may become accessible via keys in the harvested metadata.

Acceptable use of the harvesting interface

Verification of the identities in the harvesting transactions is not part of the current specifications, but it is left to individual communities to use appropriate tools (such as HTTPS, TLS) if required.

A flow control mechanism built into the harvesting protocol and error messages will give archives some level of control on the usage of their harvesting interface by service providers.

Acknowledgements

Participants at the meeting

Caroline Arms – Library of Congress
Ray Dennenberg - Library of Congress
Daniel Greenstein – Digital Library Federation
Thomas Krichel – University of Surrey
Carl Lagoze – Cornell University
Xiaoming Liu – Old Dominion University
Clifford Lynch – Coalition for Networked Information
David Millman – Columbia University
Michael L. Nelson - NASA
John Ober – University of California
Thorsten Schwander – Los Alamos Laboratory
David Stuve - MIT
Robert Tansley – University of Southampton
Hussein Suleman – Virginia Tech
Simeon Warner - Los Alamos Laboratory
Herbert Van de Sompel – Cornell University

Meeting Supported by

The Digital Library Federation
The Cornell Digital Library Group

Report by

Carl Lagoze – Cornell University - <lagoze@CS.Cornell.EDU>
Hussein Suleman – Virginia Tech - <hussein@vt.edu>
Herbert Van de Sompel – Cornell University - <herbertv@cs.cornell.edu>