

Research Metadata Harvest Project
Draft Meeting Summary, Technical Issues Meeting
Harvard University, May 1-2, 2000
DRAFT

Attending: Dale Flecker -- chair (Harvard), Caroline Arms (Library of Congress), Priscilla Caplan -- recorder (FCLA), Bernie Hurley (UC Berkeley), Carl Lagoze (Cornell), Charlene Mason (Minnesota), David Millman (Columbia), John Perkins (CIMI), John Price-Wilkin (Michigan), Thorny Staples (Virginia)

The group reviewed directions set at the March 7 meeting in Ann Arbor and the current status of the Open Archives Initiative (OAI) [1]. Technical requirements for interoperability within the OAI are defined in the Santa Fe Convention (SFC) [2]. With this background, the group focused on 1) identifying uses that could be made of harvested metadata and possible service providers, 2) reviewing the metadata requirements of the SFC for applicability to this initiative, 3) reviewing the subset of the Dienst protocol used by SFC, 4) reviewing the registry mechanism used by SFC, 5) exploring governance and support issues, and 6) identifying projects that might be proposed for funding.

1. Uses of harvested metadata

The governing assumptions were that the data harvested: would already exist (as opposed to having been created for this initiative); would be extremely heterogeneous: is frequently hidden in a wide variety of proprietary databases and is not available today for standard web harvesting. What was harvested would vary greatly from institution to institution.

The following potential uses of the metadata were proposed:

- a "super catalog" of all types of materials available to members of a consortium such as a multi-university system;
- union access to large sets of digitized materials in certain broad subject areas, e.g. Americana;
- a uniform way for an institution to share its own metadata so that it does not have to do specific programming to support individual requests;
- a preservation "registry" recording preservation decisions for items in process to help prevent duplication of effort;
- a way to make library MARC catalog holdings accessible from a general-purpose search engine;
- a way to selectively enhance a library's catalog by adding links to digital reproductions of works;
- a "universal catalog" providing access to library books, journal literature, and other textual and non-textual research materials from a single interface.

For some of these purposes, it would be useful to know whether an item is publicly available or restricted in some way, and/or whether it is electronically available or not. For others, it is necessary to know whether an item is available to a particular community (e.g. Berkeley students). Both the high utility and high difficulty of including A&I data were noted.

John Perkins explained that the museum community is less interested in scholarly research materials and more interested in K-12 and the public audience for digital tours and exhibitions. A union catalog of unconnected items is of little interest. Museums are interested in aggregated, authored information that puts materials in the context of a collection. For these applications it is necessary to link harvested discovery data with richer descriptions.

There was consensus that the greatest utility for the least effort was the sharing of digital content already freely available on the web to create cross-institutional collections of like materials. This led to the assessment of the difficulty of providing services based on various types of materials and metadata summarized below.

Material	Audience	Difficulty	Comment
A&I data	all	very hard	publishers' cooperation required
Americana	humanists, social scientists	moderate	have to identify by subject
audio collections	all	easy	
EADs	humanists	easy	already done by RLG
faculty web pages	researchers	hard	no existing metadata
GIS	researchers	hard	huge payoff
library OPACs	all	easy	minimal utility
museum collections	all	easy	need link to richer metadata
social science datasets (DDI)	researchers	medium	tie to Harvard's DLI2 project
visual materials	all	easy?	libraries own both metadata & images; will museums contribute?

2. Review of SFC metadata requirements

The SFC requires that all participating archives support the Open Archives Metadata Set (OAMS) [3]. Archives may also optionally support other metadata formats, such as Dublin Core and MARC. The protocol provides a mechanism for inquiring about what formats are available for an item, and requesting records in a particular format.

The relation of OAMS and Dublin Core 1.1 [4] was explored. Major differences include the following:

- OAMS defines nine elements of which four are mandatory; Dublin Core has fifteen elements all of which are optional.
- The textual nature of eprint archives is reflected in the OAMS, which uses the less general concepts "Author" and "Abstract" for Dublin Core's "Creator" and "Description", and does not include elements for Format or Type.
- The OAMS elements "Full Id" and "Date of Accession" are, in a metadata harvesting context, the record key and creation date of the metadata record, and as such, are administrative metadata out of scope for Dublin Core.
- The OAMS element "organization" (affiliation) was recently rejected by the Dublin Core Metadata Initiative (DCMI) as a refinement for Creator on the grounds that it does not fit the DCMI data model, in that it describes another resource (that is, "affiliation" describes the Creator, not the resource created).

The relationship between OAMS and Dublin Core 1.1 is summarized in the table below. M=Mandatory, O=Optional, R=Repeatable, NR=Not Repeatable

OAMS	Dublin Core 1.1
Title (M, NR)	Maps to Title (O, R)
Date of accession (M, NR)	No equivalent
Display ID (O, R)	Maps to Identifier (O, R)
Full ID (M, NR)	No equivalent
Author (M, R) -- contains "name" and "organization" (affiliation) elements.	"name" maps to Creator (O, R); no equivalent for "organization"
Abstract (O, NR)	Maps to Description (O, R)
Subject (O, R)	Maps to Subject (O, R)
Comment (O, R)	Possibly maps to Description (O, R)
Date for Discovery (O, R)	Maps to Date (O, R)
no equivalent	Language (O, R)
no equivalent	Publisher (O, R)
no equivalent	Contributor (O, R)
no equivalent	Relation (O, R)
no equivalent	Source (O, R)
no equivalent	Coverage (O, R)
no equivalent	Rights (O, R)
no equivalent	Format (O, R)
no equivalent	Type (O, R)

The information in Dublin Core elements without explicit equivalents in OAMS can be included in a Comment element.

Discussion centered around the following questions:

- Are we bound to adopt OAMS, or phrased differently, must (or should) the research metadata harvest initiative use the same "core" metadata element set as OAI? There was some sentiment against establishing a "competing" element set. E-print archive harvesting can be seen as a subset of the more general universe of research metadata harvesting. In this context, having multiple, mildly variant "core" sets for resource discovery could be seen as counter-productive. Also, metadata element sets require long-term maintenance, and this initiative is not necessarily equipped to be a maintenance agency.
- Is OAMS as currently defined adequate to the needs of research metadata harvesting? Some felt it should be broadened to encompass formats other than text, by generalizing some labels and including a Format element. Rights (copyright), access information (restricted/unrestricted), coverage (geographical and temporal), genre type, and citation for journal articles (journal title, volume, issue, page range) were all mentioned at some point as being potentially useful elements.
- What is the proper relationship of OAMS and Dublin Core? Dublin Core is a widely employed, *de facto* standard for course-granularity resource discovery. Many of our institutions already maintain crosswalks from MARC, EAD and other data formats to Dublin Core. The case for most OAMS deviations from Dublin Core was not fully obvious.

Given the above, the group agreed the best course of action would be to approach OAI with a request to modify OAMS, in preference to either adopting OAMS as is or developing a "core" metadata element set specifically for this initiative. Key factors in this decision included the recognition that OAMS has not been so widely implemented that modifications would be impossible, that the single OAI meeting that defined the OAMS did not have the weight of a long-term community consensus effort, and that the organizational relationship between OAI and any research metadata harvesting initiative was still undetermined.

The recommendation is to use elements from the Dublin Core namespace where there is semantic congruence, and to put unique OAMS elements in a separate OAMS namespace. OAMS "Author" is an exception; the label was broadened to "Creator" but the element remains in the OAMS namespace in order to retain the subordinate element "organization". A different subset of elements were defined as Mandatory (including Date for discovery and excluding Creator). There was much discussion of whether all Dublin Core elements should be valid (although optional) or whether only a minimal set of element should be defined. The majority opinion was to define the minimal set and add elements over time as need is demonstrated by implementers.

These revisions to the OAMS will be proposed to the OAI at their June 3 meeting in San Antonio. A revised version of the OAMS DTD is given at the end of this paper. (Note

that since XML namespace declarations do not apply to DTDs, this would be better implemented though Xschema.)

From OAMS namespace	From Dublin Core namespace
	Title (M) = OAMS Title
	Date (M) = OAMS Date for Discovery
	Description = OAMS Abstract
	Identifier = OAMS Display ID
	Subject = OAMS Subject
Creator -- contains "name" and "organization" (renamed from "Author")	
Date of accession (M)	
Full ID (M)	
Comment	

3. Review of SFC protocol requirements

The SFC specifies a subset of the Dienst protocol to be supported by data and service providers [5]. The protocol as used in the SFC assumes that all archives have unique archive identifiers, that records have identifiers which are unique within the archive, and that archives may have hierarchical partitions. Partitions are logical and may be based on any criteria upon which a subset of records may be selected from the archives.

Review of the protocol found the following issues:

- Records cannot be deleted. The ability to do this should be added to the protocol.
- There is no way for either the client or the server to control the size of a retrieval set. The ability to do this should be added to the protocol.
- The protocol has no access control functions. It was agreed that this should be done outside of the protocol. The research metadata harvesting initiative, however, should define for its own use what mode of access control (IP filtering, http BasicAuth, or certificates) participants should support.
- The protocol does not support selectivity in retrieval apart from records added after a certain date, and records from predefined partitions. The implication of this is that metadata harvesting will be broad and generic, based on data provider, date and partition only. Additional precision to support topic-, format-, or community-specific services will be implemented by the service provider by selecting from harvested metadata. The group accepted this model.

Overall the protocol was found to be adequate with requested revisions as noted.

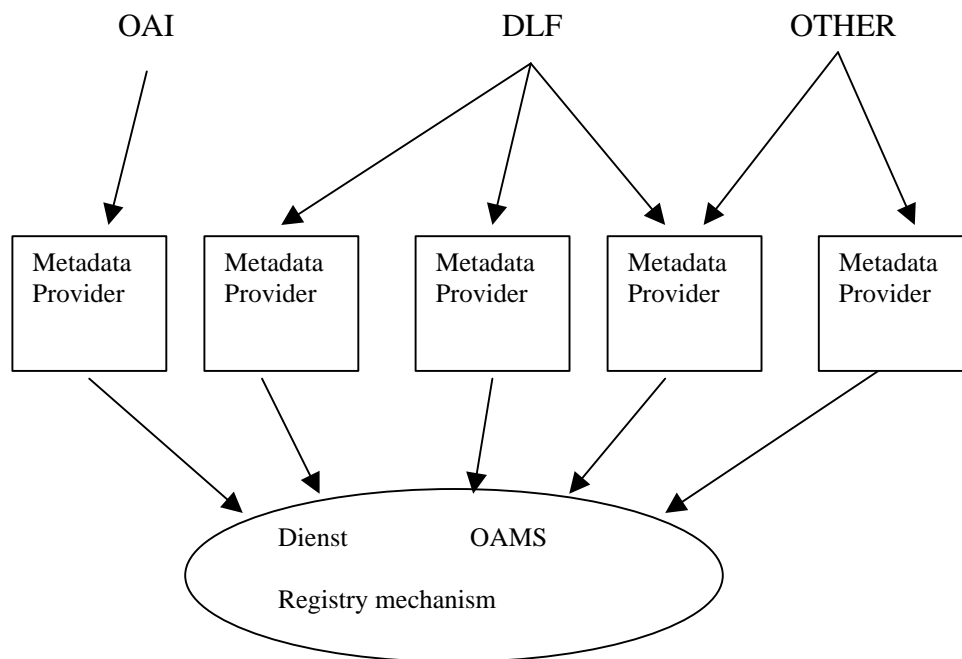
4. Review of the SFC registry mechanism

The SFC provides a template for data providers to register their participation [6]. The template and registration mechanism was found to be adequate for research metadata harvesting. A question remains as to whether there should be separate registries for different communities.

5. Governance and support

For the purpose of interoperability, the SFC metadata specification, protocol and registry should be seen as a generic technology base for open metadata harvesting. Multiple agencies can use this technology base, along with additional technology components such as access control, in the implementation of their own initiatives. In this model, OAI's eprint archive initiative and DLF's research metadata initiative are not related organizationally, but rather two separate agencies making use of the same technology base for their various projects.

Under this model, a primary governance question is who is responsible for the technology base. Some organization or institution would need to take on the role of maintenance agency, responsible for updates to standards, documentation and the metadata specification. Presumably initiatives could develop their own software, operate their own registries and define their own conventions for participation. The question of who can use the technology base was not addressed.



6. Identification of projects

Internet search engines will be contacted for their interest in a research project to develop services. Google, Excite (Jack Xal), SWISH++ and commercial intranet engines were considered most promising. Representatives from a number of search engines will be visiting Harvard later in the year, which may provide additional contacts.

Research groups focused on information retrieval (such as the one at the University of Massachusetts) will be contacted to see if they have interest in using harvested data for developing experimental services.

OCLC and RLG will be contacted for their interest in harvesting. RLG may be interested in this technology base for their Cultural Materials Initiative. RLG is a logical service provider candidate for visual materials. OCLC's CORC project should be considered a potential data provider.

Participating institutions should consider whether they could apply for grant funding to become service providers in some area. The CIC, ICPSR, and Harvard (because of their DLI-2 project) could be interested in services related to social science datasets.

REFERENCES

- [1] <http://www.openarchives.org/>
- [2] http://www.openarchives.org/sfc/sfc_entry.htm
- [3] http://www.openarchives.org/sfc/sfc_oams.htm
- [4] <http://purl.org/dc/>
- [5] <http://www.cs.cornell.edu/cdlrg/dienst/protocols/OpenArchivesDienst.htm>
- [6] http://www.openarchives.org/sfc/data_provider_template.htm

PROPOSED REVISED OAMS DTD

```
<!-- Open Archives Metadata Set (oams) -->
<!-- revision proposed by research metadata harvesting initiative -->
<!-- May 2000 -->
<!-- Dates are to be encoded using the "Complete Date" variant of
ISO8601-->
<!ENTITY % doctype "oams">
<!ELEMENT %doctype; (dc:title, dc:date+, dc:description*, dc:identifier*, dc:subject*,
oams:creator*, oams:accession, oams:fullId, oams:comment*)>
<!ELEMENT dc:title (#PCDATA)>
<!ATTLIST dc:title xmlns:dc CDATA #FIXED "http://purl.org/dc">
<!ELEMENT dc:date (#PCDATA)>
<!ATTLIST dc:date xmlns:dc CDATA #FIXED "http://purl.org/dc">
```

```

<!ELEMENT dc:description (#PCDATA)>
<!ATTLIST dc:description xmlns:dc CDATA #FIXED "http://purl.org/dc">
<!ELEMENT dc:identifier (#PCDATA)>
<!ATTLIST dc:identifier xmlns:dc CDATA #FIXED "http://purl.org/dc">
<!ELEMENT dc:subject (#PCDATA)>
<!ATTLIST dc:subject xmlns:dc CDATA #FIXED "http://purl.org/dc">
<!ELEMENT oams:creator (name, organization*)>
<!ATTLIST oams:creator xmlns:oams CDATA #FIXED
"http://www.openarchives.org/sfc/sfc_oams/htm">
<!ELEMENT oams:name (#PCDATA)>
<!ELEMENT oams:organization (#PCDATA)>
<!ELEMENT oams:accession EMPTY>
<!ATTLIST oams:accession
xmlns:oams CDATA #FIXED "http://www.openarchives.org/sfc/sfc_oams/htm"
date CDATA #REQUIRED>
<!ELEMENT oams:fullId (#PCDATA)>
<!ATTLIST oams:fullId xmlns:oams CDATA #FIXED
"http://www.openarchives.org/sfc/sfc_oams/htm">
<!ELEMENT oams:comment (#PCDATA)>
<!ATTLIST oams:comment xmlns:oams CDATA #FIXED
"http://www.openarchives.org/sfc/sfc_oams/htm">
<!-- ENTITY sets - from MathML DTD go here -->
<!-- one might also consider declaring a "math" namespace for these -->

```