Diversity in Metadata Schemes used by OAI-PMH Data Providers

Sanjeev K Sunny

Central University of Bihar
B.I.T. Campus, P.O. – B.V. College,
Patna – 800 014
sunny@cub.ac.in

Abstract.‘Digital Repositories or Archives’ or ‘Digital Libraries’ use a set of elements to describe the characteristics of each document. This set of elements is called metadata schema. Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) evolved as a means to achieve Interoperability among repositories. OAI-PMH mandates the oai_dc schema (based on unqualified Dublin Core) as a minimum standard for interoperability (Lowest Common Denominator).oai_dc is a simple format providing baseline interoperability. It may not be suitable for every repository, service or community to share only oai_dc. Many digital repositories have developed other metadata schemes, as per their specific needs, by extending oai_dc or with completely different set of elements. The author intends to identify all the metadata schemes being used by open access digital repositories with their popularity in terms of instances of their use.

Keywords: Digital Library, Metadata, Metadata Harvesting

1. Introduction

There are hundreds of digital repositories (archives) which support and participate in the Open Archives Initiative (OAI). These archives support Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). Supporting OAI-PMH warrants the availability of metadata records of the resources for harvesting and it mandates the ***oai_dc*** schema (based on unqualified Dublin Core) as a minimum standard for interoperability (Lowest Common Denominator). oai_dc is a simple format providing baseline interoperability. It may not be suitable for every repository, service or community to share only oai_dc. The 15 Dublin Core Metadata Element Set (DCMES) may not include enough of the elements required by different repositories. In this case one can create a new schema incorporating the additional required elements alongwith those in DCMES. The elements of oai_dc may not be sufficiently precise for one's metadata records, as DCMES is an 'unqualified' metadata encoding schema. In this case one can get greater precision by creating a new schema adding 'encoding schemes' to existing DCMES elements. DC may not be the metadata format required by every repository. In a particular community one may want to exchange metadata in another format, for example, in IMS/IEEE LOM for e-Learning metadata or in ODRL (Open Digital Rights Language).For one or other reasons many metadata schemes have been developed and being used by repositories.

This paper identifies the metadata schemes being used by open access digital repositories with their popularity in terms of their instances of use. This paper explains how the open access digital repositories were identified for this study and how the metadata schemes were collected. The findings of the study i.e. the popularity of metadata schemes among open access digital repositories in terms of instances of their usage are given in the last section.

2. Background

Lots of research and educational material are produced by members of a research university or organization in digital format, much of which are never published by traditional means. It is essential to protect the significant scholarly assets of the institution as their constituents produce increasing amounts of original material in digital formats called e-prints (referred as 'documents' here onward). Repository of these documents are called 'institutional repositories', 'e-print repositories' or 'digital repositories or archives' or ''digital libraries'. These repositories use a set of elements to describe the characteristics of each document. This set of elements is called *metadata schema*. Interoperability among repositories could be achieved by sharing these metadata but initially there was no machine-based way of sharing these metadata. In order to find a way of interoperability among these repositories a meeting was called in October 1999 at Santa Fe, New Mexico. Two possible approaches were identified to achieve interoperability: cross-searching multiple archives based on a protocol such as Z39.50 or else harvesting their metadata at one place and make it accessible from single interface. The later approach resulted in what we know today as Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH).

OAI-PMH defines a mechanism for harvesting records of documents containing metadata from repositories. OAI-PMH defines two logical roles: “Data Provider” and “Service Provider”. A *Data Provider* maintains one or more repositories (web servers) that support the OAI-PMH as a means of exposing metadata. They are the creators and keepers of the metadata and repositories of resources. A *Service Provider* issues OAI-PMH requests to data providers and uses the metadata as a basis for building value-added services. A Service Provider in this manner is "harvesting" the metadata exposed by Data Providers. They use the harvested metadata for the purpose of providing one or more services across all the data. OAI-PMH uses XML Schema to define record formats. Data providers can exchange any metadata using OAI-PMH as long as it can be encoded as XML and an XML Schema is defined for it. OAI-PMH mandates the ***oai_dc*** schema (based on unqualified Dublin Core) as a minimum standard for interoperability (Lowest Common Denominator). It defines a container schema that is OAI-specific, and is hosted on the OAI Web site. It imports a generic DCMES (DC Metadata Element Set) schema. The generic DCMES schema is hosted on the DCMI (Dublin Core Metadata Initiative) Web site.

3. Scope and Objectives

The mandated schema oai_dc is a simple format providing baseline interoperability; so, it may not be suitable for every repository or service. There are many other metadata schemes, both extended from oai_dc and with completely different sets of elements, being used by data providers in order to achieve maximum accessibility. The objective of this research was to explore the entire metadata schemes being used by OAI-PMH data providers. The author intended to find out the number of metadata schemes being used by OAI-PMH data providers and the frequency of their uses. Many of these schemes are merely extension of oai_dc and many have completely different sets of elements. So, it was also aimed to find how many of them are merely extension of oai_dc and how many of them are completely different from that of oai_dc. Thorough study of every schema is out of the scope of this paper.

4. Methodology

Supporting OAI-PMH warrants the availability of metadata records of the resources for harvesting. These metadata records should be network-accessible. Every archive provides a web-accessible URL which accepts the OAI-PMH requests. This URL is called “base URL” of the repository. This section lists various sources which have been used to collect base URLs of data providers; the OAI-PMH verb ‘ListMetadataFormats’ used with every base URL to get the name and location of metadata schema being used. After collecting names and locations of base URLs each schema has been studied to find out its frequency of usage and to know whether they are extended from oai_dc or they have completely different set of elements.

In order to get the name and location of metadata schema being used by these repositories, their base URLs were required. Open Archive Initiative maintains a list of registered OAI conforming repositories at its site. There are some other sources available on web which provides list of repositories with their base URLs. Following is a list of resources which were used to collect baseURLs for this research:

- *Registered data providers*

  This list is maintained by Open Archives Initiative. This list can be found at “http://www.openarchives.org/Register/BrowseSites”.

- *Open Language Archives Community (OLAC)*

  A machine readable list of registered archives is available at “http://www.language-archives.org/register/archive_list.php4”.

- *Celestial – Registered Archives*

  A list of 1014 archives registered with Celestial is available at “http://celestial.eprints.org/”.

- *Registry of Open Access Repositories (ROAR)*

  1113 archives are registered with ROAR, which is available at “http://roar.eprints.org/”.

These lists provide names of the repositories along with their base URLs. All the above lists are available either in HTML format or in XML. These HTML and XML files have been used to extract the base URLs of the repositories. There were repetitions of the base URLs among different lists i.e. many base URLs were present in more than one lists. A list of unique base URLs was extracted. This list contains 1992 unique base URLs which have been used to study the metadata schemes.

As stated above, it is the base URL of the repository which accepts, processes and responds to the OAI-PMH requests (verbs). The OAI-PMH verb "ListMetadataFormats" (described in the next section) has been issued to all the above 1992 base URLs. Against 1992 requests there were only 1471 responses. Others resulted in many types of HTTP errors as mentioned in the following section. Out of 1471 responses, 62 responses were without metadataPrefix i.e. only 1409 responses provided the list of metadata formats being used by them. All the research findings given in section 5 are based on these 1409 responses.

ListMetadataFormats verb was used with base URLs of all the 1992 repositories and 1408 repositories responded with the metadataPrefix and schema locations. 182 metadataPrefixes from all these responses have been extracted. Similarly the entire schema locations have been extracted. These schemes have different levels of popularity (in terms of their usage) among the data providers. Some are very popular and are being used by many data providers, while some are being used by only one data provider.

4.1 ListMetadataFormats : The OAI-PMH Verb

This verb is used to retrieve the metadata formats available from a repository. An optional argument "identifier" restricts the request to the formats available for a specific item. It is an *optional* argument that specifies the unique identifier of the item for which available metadata formats are being requested. If this argument is omitted, then the response includes all metadata formats supported by this repository. Note that the fact that a metadata format is supported by a repository does *not* mean that it can be disseminated from all items in the repository. For example

Request

To list the metadata formats that can be disseminated from the repository `http://memory.loc.gov/cgi-bin/oai`following request is given

```
http://memory.loc.gov/cgi-
bin/oai?verb=ListMetadataFormats
```

Response
Here is the response of the above request

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
         http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
<responseDate>2002-06-08T15:19:13Z</responseDate>
<request verb="ListMetadataFormats">
           http://memory.loc.gov/cgi-bin/oai</request>
<ListMetadataFormats>

<metadataFormat>
<metadataPrefix>oai_dc</metadataPrefix>
<schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</sche
ma>
<metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_d
c/
</metadataNamespace>
</metadataFormat>

<metadataFormat>
<metadataPrefix>oai_marc</metadataPrefix>
<schema>http://www.openarchives.org/OAI/1.1/oai_marc.xsd</sc
hema>
<metadataNamespace>http://www.openarchives.org/OAI/1.1/oai_m
arc
</metadataNamespace>
</metadataFormat>

</ListMetadataFormats>
</OAI-PMH>
```

The response shows that the repository supports two metadata formats:`oai_dc`, and `oai_marc`(Text shown in bold). For each of the formats, the metadataPrefix used for the schema is given within `<metadataPrefix>…</metadataPrefix>`and the location of an XML Schema describing the format is given within `<schema>..</schema>` tags.

4.2 Error messages

As stated above, only 1471 out of 1992 repositories responded to the OAI-PMH verb request with metadataPrefix and schema locations. Other 521 repositories could not be accessed; HTTP error messages instead of OAI-PMH response were received. The list of HTTP error messages received is given below:

301: Moved Permanently
302: Moved Temporarily
400: Bad Request
401: Unauthorized
403: Forbidden
404: Not Found
501: Not Implemented
502: Bad Gateway
503: Service Unavailable
504: Gateway Time-out

5. Findings

As stated in section 3.2, 1408 repositories responded with "metadataPrefix" tag and the schemes have different levels of popularity among the data providers. When these schemes have been studied some of these schemes have been found to be simply extended from oai_dc and others with completely different set of elements. Though these schemes are being classified on the basis of above information i.e. whether they are extended form oai_dc or not, one more group has been created on the basis of the popularity. Thus the metadata schemes have been divided into three groups – "Most widely admired schemes", "Schemes extended from oai_dc" and "Schemes with completely different set of elements". These are mentioned in the following sections.

5.1 Most widely admired schemes

The schemes which have been most widely admired by OAI-PMH data providers are given in Table 1 (in decreasing order of their instances of use). This table consists of schemes extended from oai_dc as well as schemes with completely different set of elements. MARC21slim, rfc1807, oai_marc, mods, mets and didmodel are different from oai_dc. These schemes have completely different set of elements. While etdms, uketd_dc, context_object and qdc are extended from oai_dc. It means these schemes have elements from oai_dc as well as defined by them and/or from some other schemes.

S.No.	Metadata Schema	Instances

1	oai_dc	1354
2	MARC21slim	199
3	rfc1807	146
4	oai_marc	140
5	mods *(three versions)*	125
6	etdms	92
7	mets	80
8	didmodel	70
9	uketd_dc	68
10	context_object	64
11	qdc	49

Table 1: Most widely admired schemes

5.2 Schemes extended from oai_dc

Table 1 includes 4 most popular schemes which have been extended from on oai_dc. Other schemes which are extended from oai_dc have been divided into three groups – one having at least 10 instances of usage, another having less than 10 instances but more than one instance and the third with single instance. Table 2.1 enlists the first group of schemes i.e. with at least 10 instances of usage.

S.No.	Metadata Schema	Instances
1	junii2	46
2	xmetadiss	37
3	junii	32
4	PROPRINT_METADATA_SET	28
5	attribute-schema	21
6	olac	21
7	xmetadissplus	20
8	dcterms	13
9	amf	13
10	nsdl_dc*(Three Versions)*	25

Table 2.1: oai_dc based schemes with at least 10 instances

One thing is important to note that “junni2” and “junii” are two versions of the same schema so, cumulative instance of its usage becomes 78. Similarly, “xmetadiss” and “xmetadissplus” are two versions of the same schema so, its cumulative instances of usage becomes 57. “nsdl_dc” has three versions but none of those have significant instances so, it has been placed at single place.

Table 2.2 enlists second group of schemes i.e. less than 10 instances but more than one instance of usage.

S.No.	Metadata Schema	Instances
1	aofr	7
2	nereusx	5
3	epdcx	5
4	bibliographic	3
5	pndsdc	2
6	pa	2
7	ems	2
8	dc_citation	2
9	dc-schema	2
10	CICQualifiedDC	2

Table 2.2: oai_dc based schemes with 2 to 9 instances

Besides the oai_dc based schemes given in Table 2.1 and Table 2.2, there are many schemes which have single instance of usage. These are given in Table 2.3.

S.No.	Metadata Schema	Instances
1	ads_dc	1
2	agris_ap	1
3	collexis	1
4	dare_qdc	1
5	iesr	1
6	imlsdccprofile	1
7	kmoddl_v1.00	1
8	oai_qdc	1
9	object	1
10	oszkint	1
11	oszkqdc	1
12	picture	1

13	rdn_dc	1
14	schema_uc	1
15	yale_dc	1

Table 2.3: oai_dc based schemes with single instance

5.3 Schemes with completely different set of elements

Table 1 includes 6 most popular schemes with completely different set of elements. There are many more such schemes. Other schemes with completely different set of elements have been divided in to three groups – one having at least 10 instances of usage, another having less than 10 instances but more than one instance and the third with single instance. Table 3.1 enlists the first group of schemes i.e. with at least 10 instances of usage.

S.No.	Metadata Schema	Instances
1	xepicur	44
2	didl	29
3	arno	10

Table 3.1: Schemes with different set of elements having at least 10 instances

Table 3.2 which follows, enlists second group of schemes i.e. less than 10 but more than one instance of usage.

S.No.	Metadata Schema	Instances
1	Datatype-en	9
2	DIDL	6
3	Archivearticle	6
4	Rdf	5
5	Mtd-br	5
6	VOResource	5
7	arXive	4
8	arXiveOld	4
9	arXiveRaw	4
10	Zim_export	3
11	Tel	3
12	Inria	3
13	Akf	3

14	Zthes	2
15	Xhtml-transitional	2
16	Rugdb	2
17	Record	2
18	Hal	2
19	Lom	2
20	Mabxml	2
21	Gmd	2
22	Dited	2
23	Cnr_eprints	2
24	Brief-record	2
25	Asic	2
26	XMLSchema	2
27	Unimarc	2
28	DDF_MXD_Schema	2

Table 3.2: Schemes with different set of elements having 2 to 9 instances

Table 3.1 and Table 3.2 included those schemes which have completely different set of elements and have at least 2 instances of usage. There are many schemes which have single instance of usage. These are given in Table 3.3.

S.No.	Metadata Schema	Instances
1	annotation	1
2	article	1
3	ben	1
4	bibl	1
5	ccsd_mem	1
6	cstc	1
7	dif_v9.4	1
8	dif_v9.7	1
9	doajArticle	1
10	dsOverview	1
11	eruditarticle	1
12	IMDI_3.0	1

13	ims	1
14	imsmd_v1p2p2	1
15	imsmd_v1p2p4	1
16	info-uri-registry	1
17	MetaData	1
18	Monograph	1
19	mtd2-br	1
20	native_xml	1
21	news-opps	1
22	olac-archive	1
23	Periodical	1
24	Version2-0 DDI	1
25	xlink	1
26	xrefer	1

Table 3.3: Schemes with different set of elements having single instance

6. Conclusion

There is immense diversity in the types of information objects. Describing these varied information objects using any single metadata schema is not feasible. Presently, Open Archives Initiative mandates oai_dc as a minimum standard for interoperability (Lowest Common Denominator). The varied nature of metadata schemes can be easily seen. By looking on the diversity of metadata schemes being used by OAI-PMH data providers, it is obvious that oai_dc is not sufficient for every information objects.

After oai_dc, XML schema of MARC21 i.e. “MARC21slim” is second most popular schema (being used by 199 repositories). There is one more XML schema for MARC21, “oai_marc” which was created before MARC21slim. Though, usage of MARC21 instead of oai_marc is strongly recommended since the release of the XML Schema for MARC21 metadata by the Library of Congress, June 2002, it is being used by 140 repositories. It is recommended that the metadataPrefix "marc21" be used with this metadata format. Rfc1807 is very simple schema and is being used by 146 data providers. Besides above three schemes other popular schemes are MODS, METS and DIDMODEL. The number of metadata schemes created by extending oai_dc is very large. ETDMS (92), UKETD_DC (68) context_object (64) and Qualified Dublin Core (49) are most popular metadata schemes based on oai_dc. Other popular metadata schemes based on oai_dc are XMetaDiss (37), junii (32), PROPRINT_METADTA_SET (28), attribute-schema (21), OLAC (21), etc. There are many such metadata schemes which are being used by only one repository. These schemes are of both types i.e. extended from oai_dc as well as completely different from oai_dc. Single instances of many metadata schemes prove that there is a huge demand of different types of metadata schemes by various repositories in order to describe their information resources in an efficient manner.

7. Reference

Carl Lagoze, Herbert Van de Sompel, Michael Nelson & Simeon Warner (Eds) (2002). *An XML Schema to represent MARC records*, Retrieved July 15, 2011from Open Archive Initiative at http://www.openarchives.org/OAI/2.0/guidelines-oai_marc.htm.

Carl Lagoze, Herbert Van de Sompel, Michael Nelson & Simeon Warner (Eds) (2004). *The Open Archives Initiative Protocol for Metadata Harvesting*, Retrieved July 15, 2011from Open Archive Initiative at http://www.openarchives.org/OAI/openarchivesprotocol.html

Carl Lagoze, Herbert Van de Sompel, Michael Nelson & Simeon Warner (Eds) (2005). *Conveying rights expressions about metadata in the OAI-PMH framework*, Retrieved July 15, 2011from Open Archive Initiative at http://www.openarchives.org/OAI/2.0/guidelines-rights.htm.

MARCXML Official Website.Retrieved July 15, 2011 from http://www.loc.gov/standards/marcxml/.

Open Archives Forum (a). *OAI for Beginners - the Open Archives Forum online tutorial*.Retrieved July 15, 2011 from Open Archives Forum athttp://www.oaforum.org/tutorial/.

Open Archives Forum (b). *Experiences - Implementation Tests: OAI Service Provider*, Retrieved July 15, 2011 from Open Archives Forum athttp://www.oaforum.org/resources/tvserviceimpl.php.

Open Archives Initiative (2000). *Schema for rfc1807 metadata format,*Retrieved July 15, 2011 from Open Archives Initiative at http://www.openarchives.org/OAI/1.1/rfc1807.xsd.