ISSN (Print): 0972-2467
ISSN (Online): 0976-2477
SRELS Journal of Information Management, Vol 55(1), February 2018, p. 1-7 DOI: 10.17821/srels/2018/v55i1/116603

Harvesting of Additional Metadata Schema into DSpace through OAI-PMH: Issues and Challenges

Anup Das* **and B. Sutradhar**

Central Library, Indian Institute of Technology (IIT) Kharagpur – 721302, West Bengal, India;
anupdas1704@gmail.com, bsutra@library.iitkgp.ernet.in

Abstract

The focus of this paper is to present a manual for harvesting added metadata schema elements from data provider repository through PAI-PMH. Here we have attempted to extend the ability of DSpace as well as the OAI-PMH to host or share other metadata schema elements e.g. LRMI, IEEE-LOM, ETD etc. DSpace, by default, is capable of importing and exporting only Dublin Core metadata; but DSpace has the capability by which one can define his own submission form and expose customized fields to OAI-PMH request. This paper attempts to provide a guideline to a DSpace administrator to accommodate different metadata schema as well as enhancing the Interoperable capability between Data provider and Service provider repository for Importing and Exporting customized metadata Schema. This is an important concern as Institutions have different types of content and have implemented metadata schema and elements appropriate to their content.

Keywords: Data Provider, DSpace Intermediate Metadata (DIM), Dublin Core Metadata Initiative (DCMI), IEEE-LOM, Learning Resource Metadata Initiative (LRMI), Metadata Harvesting, Qualified Dublin Core (QDC), Service Provider, MPEG-7, OAI-PMH

1. Introduction

Digital India is a campaign initiated by the Government of India in 2015 to ensure that Government services are made available to every citizen electronically by improved online infrastructure and by increasing Internet connectivity. The vision of Digital India programme is inclusive growth in areas of electronic services, products, manufacturing and job opportunities, etc. and it is centered on three key areas – Digital Infrastructure as a Utility to Every Citizen, Governance and Services on Demand and Digital Empowerment of Citizens.

In the age of information explosion and the rapid growth of digitization activities, number of digital repositories of various educational and research institutions situated in India and abroad are increasing. But no single initiative can assimilate all the research output into one umbrella. Information is scattered and it is very difficult for new learners or a research scholar to find relevant sources. In this situation no one can ensure that digitization will solve the problem of information access. A significant issue with digital repositories is the lack of interoperability. If two or more systems are capable of communicating with each other, they exhibit syntactic interoperability when using specified data formats and communication protocol. According to Priscilla Caplan search interoperability is 'the ability to perform a search over diverse set of metadata records and obtain meaningful results'. Interoperability touches many diverse aspects including metadata standards, underlying architecture, openness to the creation of third-party digital library services, integration with the established mechanism of scholarly communication, usability in a cross-disciplinary context. In the NDL we were facing a problem and challenge regarding various metadata standard harvesting through OAI-PMH as- Qualified Dublin Core (QDC), IEEE-LOM, Learning Resource Metadata Initiative (LRMI), MPEG-7 etc. We have overcome this problem

2. OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting (referred to as the OAI-PMH in the remainder of this document) provides an application-independent interoperability framework based on metadata harvesting.

*Author for correspondence

There are two classes of participants in the OAI-PMH framework:

- Data Providers administer systems that support the OAI-PMH as a means of exposing metadata, and
- Service Providers use metadata harvested via the OAI-PMH as a basis for building value-added service.

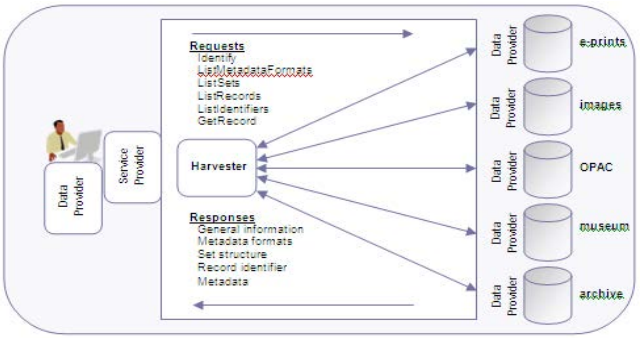

Figure 1. OAI-PMH Structure.

The OAI PMH has six request syntaxes that are used to send request to the data providers. These are:

- Identify,
- ListMetadataFormats,
- ListSets,
- GetRecord,
- ListIdentifiers, and
- ListRecords.

3. Understanding Metadata

Metadata is defined as "data about data" and as "data [information] that provides information about other data". There are three distinct types of metadata as: 1. Descriptive metadata, 2. Structural metadata, and 3. Administrative metadata.

3.1 Qualified Dublin Core

The Dublin Core Metadata Initiative (DCMI) set is a vocabulary of fifteen elements for use in resource description. The Dublin Core Schema is a small set of vocabulary terms that can be used to describe web resources (video, images, web pages, etc.), as well as physical resources such as books or CDs, and objects like artworks. The full set of Dublin Core metadata terms can be found on the DCMI website.

3.2 LRMI Metadata

The LRMI is a project led by Creative Commons (CC) and the Association of Educational Publishers (AEP) to establish a common vocabulary for describing learning resources. The LRMI project was initiated in July 2011 to make it easier for teachers and learners to find educational materials through major search engines and specialized resources discovery services (Barker and Campbell, 2014). To understand the approaches of LRMI, schema.org has been introduced.

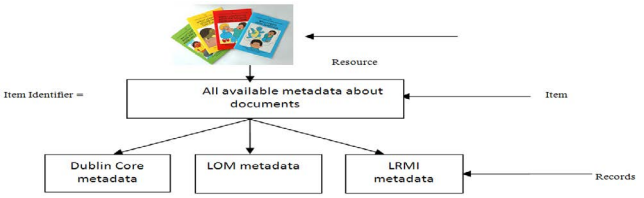

Figure 2. Metadata Structure.

3.3 IEEE-LOM

Learning Object Metadata, usually encoded in xml, is used to describe a learning object and similar digital resources used to support learning. The purpose of learning object metadata is to support the responsibility of learning objects, to aid discoverability, and to facilitate their interoperability, in the context of online Learning Management System (LMS). The IEEE 1484.12.1-2002 Standard for Learning Object Metadata (LOM) is an internationally recognized open standard for the description of "learning objects".

4. Harvesting

Harvesting is a process of gathering/assimilating together metadata from separate digital repositories into a service provider repository.

5. Harvesting of Customized Metadata Schema

Most digital library software like DSpace are metadata-schema-independent, so that the DSpace administrators may adopt customized metadata scheme most appropriate to the repository or a collection in the repository for Example: IEE-LOM, LRMI etc. Presently DSpace follows Qualified Dublin Core for input and exposes Unqualified Dublin Core through OAI-PMH protocol LRMI metadata fields in DSpace.

Table 1. DC Metadata Registry

Elements	Qualifier	Fields	Scope Note
contributor	advisor	dc.contributor.advisor	Use primarily for thesis advisor.
contributor	author	dc.contributor.author	Author(s) of the work (used by default)
title		dc.title	Varying form of title proper appearing in item.
date		dc.date	Date of publication or distribution.

To accommodate customized metadata schema or metadata fields other than Dublin Core like LRMI and IEEE-LOM following steps are to address.

- Adding new metadata elements and schema,
- New customized Input Forms for item submission,
- Indexing, and
- Import/Export.

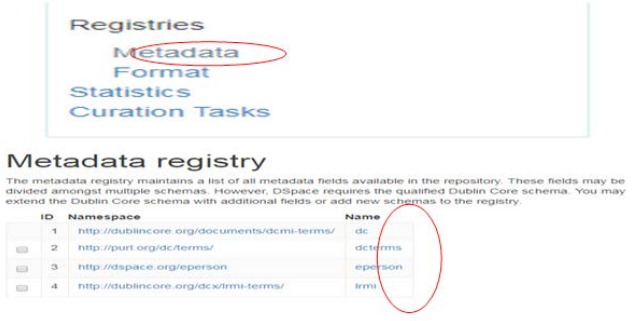

Figure 3. New Metadata Schema Registry.

5.1 New Customized Input Forms for Item Submission

As per previous discussion the default DSpace Dublin Core Metadata Registry was originally derived from the 15 Dublin Core elements. This registry initializes the default schema, where dc is used to identify the namespace. There are three types of context which DSpace administrator can add according to his/her own needs and base of knowledge regarding metadata. Following are the context:

- Default Dublin Core (DC) Metadata Registry,
- Dublin Core Terms Registry (DCTERMS), and
- Default Bitstream Format Registry.

5.1.1 Default Dublin (DC) Core Metadata Registry

As per previous discussion the default DSpace Dublin Core Metadata Registry was originally derived from the 15 Dublin Core elements. This registry initializes the default schema, where dc is used to identify the namespace.

DSpace administrator can add multiple metadata fields into existing schema like "dc" Name and Namespace or "dcterms" based on the contents which an institution have. Administrator can update, delete and move to another schema.

5.1.2 Dublin Core Terms Registry (DCTERMS)

This registry initializes an optional metadata schema, where 'dcterms' is used to identify the namespace. The registry and schema were added as a first step to facilitate the future migration of the DSpace specific DC schema, to this schema that complies with current Dublin Core standards. The main advantage of the DCTERMS schema is that no field name details gets lost during harvesting, as opposed to harvesting of so called "simple" Dublin Core, where the qualifiers from the above schema are omitted during harvesting.

Figure 4. New Metadata Field Registry.

Table 2. DC Terms Registry

Elements	Fields	Scope Note
Table of Contents	dc. Table of-Contents	A list of subunits of the resource.
abstract	dc. abstract	A summary of the re-source.
type	dc.type	The nature or genre of the resource.
subject	dc. subject	The topic of the resource.

5.1.3 Default Bitstream Format Registry

In Item submission default bitstream format metadata registry, administrator can include various "mimetype" and file "extensions". Example: Microsoft Word, Adobe PDF, SGML, Microsoft Excel, Microsoft PowerPoint,

MPEG Audio, image/PNG, XML, MPEG.

5.1.4 Adding New Metadata Schema

Administrator can also add multiple metadata schemas to describe different objects in multiple ways. Just fill the "namespace" which should be an established URI location for the new schema for further reference. "Name" should be shorthand notation for the schema this will be used to prefix a field's name (e.g. "dc", "lrmi" and "ieee-lom"). The "name" should be less than 32 characters and cannot include spaces, periods or underscore.

Figure 5. Added Schema.

Figure 6. Added Metadata Field.

5.2 New Customized Input Forms for Item Submission

Tags in 'input-forms.xml'
input-forms.xml

For clear understanding of the notation, go to DSpace installation directory. It could be any directory like "home/dspace/IDR/dspace_inst", "/home/dspace/dspace", "/home/dspace", "/usr/local/dspace", "/dspace" or"/opt/dspace". For example, if you come across $DSPACE_HOME/config or /home/dspace/IDR/dspace_inst/config, it indicates "config" directory under dspace home directory where DSpace is installed.

If Dspace administrator wants to create a separate metadata input form for different items submission using LRMI and LOM or both together, one has to modify the /home/dspace/IDR/dspace_inst/config/input-forms.xml file. The file "input-forms.xml" should be modified through this way:

LRMI Implementation

An input forms starts with '<input-forms>' and end with '</input-forms>' under this tag all the fields, value pair, showed value, store value and free text instruction are there. e.g.

```
<input-forms>
  <form-map>
        <name-map collection-handle="default" form-name="traditional" />
  </form-map>
   <form-definitions>
     <form name="traditional">
                <!-- common metadata fields -->

       <page number="1">
                <field>
              <lrmi-schema>lrmi</lrmi-schema><lrmi-element>typicalAgeRange</lrmi-element><lrmi-qualifier></lrmi-qualifier>
       <repeatable>true</repeatable><repeatable>false</repeatable>
              <label>Typical Age Range</label>
             <input-type value-pairs-name="domain_lrmi_typicalAgeRange">list</input-type>
             <hint>Select the age range(s) of user, for whom the content is targeted</hint>
          </field>
<field>
```

Each start <field> and </field> tag contains one metadata field or elements by which DSpace administrator can define various attributes like its type, what is the name, whether the field is repeatable, heading, any help message, whether the field is mandatory or optional etc.

<dc-schema>

This tag is used to define a schema. As DSpace supports Dublin Core 'dc' but Dspace administrator can add here LOM or LRMI e.g.

<dc-schema>dc</dc-schema> or <lrmi-schema>lrmi</lrmi-schema> or <lom-schema>lom</lom-schema>

<dc-element>

This tag is used to define a metadata field or element in a particular schema like 'dc', 'lrmi' and 'lom'. But DSpace by default using Dublin Core metadata schema here we can see it as 'dc-element'. DSpace administrator can define new element according his requirements as 'lrmi-element', 'lom-element' and 'etd-element' etc. e.g.

<dc-element>title</dc-element> or <lrmi-element>educationalUse</lrmi-element> or <lom-element>entity</lom-element>

<dc-qualifier>

This means a qualifier of an element. e.g.

<lrmi-element>educationalAlignment</lrmi-element> <lrmi-qualifier>educationalFramework</lrmi-qualifier>

<repeatable>

It may be either true or false. If the field is repeatable like subject or title, it should be true. If not the value should be false e.g.

<repeatable>true</repeatable> or <repeatable>false</repeatable>

There are another tags like <label>, <input-type... for mentioning one box or two box or dropdown, <hint> means for help message.

<required>

This tag is used to indicate that the field is mandatory or not. If you create this field as mandatory, can't skip or go forward without any input e.g.

<required>Enter the Source URI of the Institution</required>

Lastly there are other tags namely- '<form-value-pairs>', '<value-pairs…', '<pair>', '<displayed-value> and '<stored-value>' e.g.

```
<value-pairs value-pairs-name="domain_dc_description_searchVisibility" dc-term="dc_description_searchVisibility">
    <pair>
  <displayed-value>True</displayed-value>
  <stored-value>true</stored-value>
    </pair>
    <pair>
  <displayed-value>False</displayed-value>
  <stored-value>false</stored-value>
        </pair>
      </value-pairs>
```

5.3 Indexing

It is not necessary to change the indexing pattern but one can change index for search result. Another very crucial point is that for OAI metadata indexing, DSpace is using solar for indexing. DSpace solar dose not index metadata automatically which has been imported through manual submission process or batch import using SIP or AIP. So Data Provider Repository (DPR) administrator has a crucial role to overcome this issue, Export metadata through OAI. The kind of error that may appear in OAI request is:

http://cwds.ndl.iitkgp.ac.in/oai/request?verb=ListRecords&metadataPrefix=oai_dc&set=col_123456789_14

It is showing that no item is indexed. "Error No matches for the query". There is a command which DPR administrator has to issue so as to index and resolve the problem; Go to Dspace installation folder through terminal the path will be the following:

Type this into terminal

IDR$ cd /home/dspace/IDR/dspace_inst/bin

Then type

Dspace_inst/bin$./dspace oai import –o –c

5.4 Import/Export

5.4.1 Harvesting LRMI metadata fields in DSpace

The scenarios are many – an educator compiling a syllabus or lesson plan, a student looking for a piece of information to complete a homework assignment, or a Ph.D. candidate conducting research for their thesis – but the outcome is often the same. In the current landscape of online search, individuals seeking educational content are spending far too much time sorting through pages and pages of content without satisfactory results. This is the main problem LRMI seeks to address. If all educational content – or at least a critical mass – is described in a consistent and uniform manner, web searches will return more relevant results, and it will become easier for search engines to provide filters by which users can narrow down results even more precisely.

LRMI seeks to define the parameters by which users might search and filter learning resources online – in other words, to decide what shows up when you do a search for "math book." Once the LRMI framework is fully implemented by DSpace administrator and by search engines, users will be able to narrow search results according to terms such as subject area, age range, type of resource, alignment to educational standards, etc.

5.4.2 Challenges

While harvesting metadata from external source/ repository or data provider using DSpace, latest version provides three standard options: 1. Simple DC 2. Qualified Dublin Core (QDC), and 3. DSpace Intermediate Metadata (DIM) from harvester side. From NDL perspective --- Using SIP and AIP ingestion which involves Bulk metadata upload that also includes LRMI fields along with Bitstream** & another through custom"input-forms. xml" submission are the two ways to directly input data to

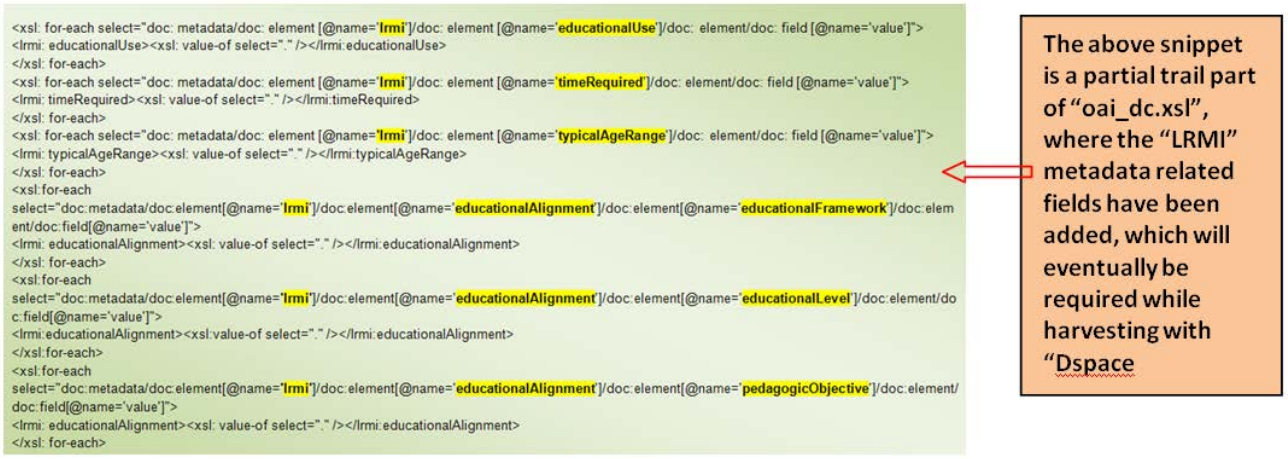

Figure 7. oai_dc.xsl Modification.

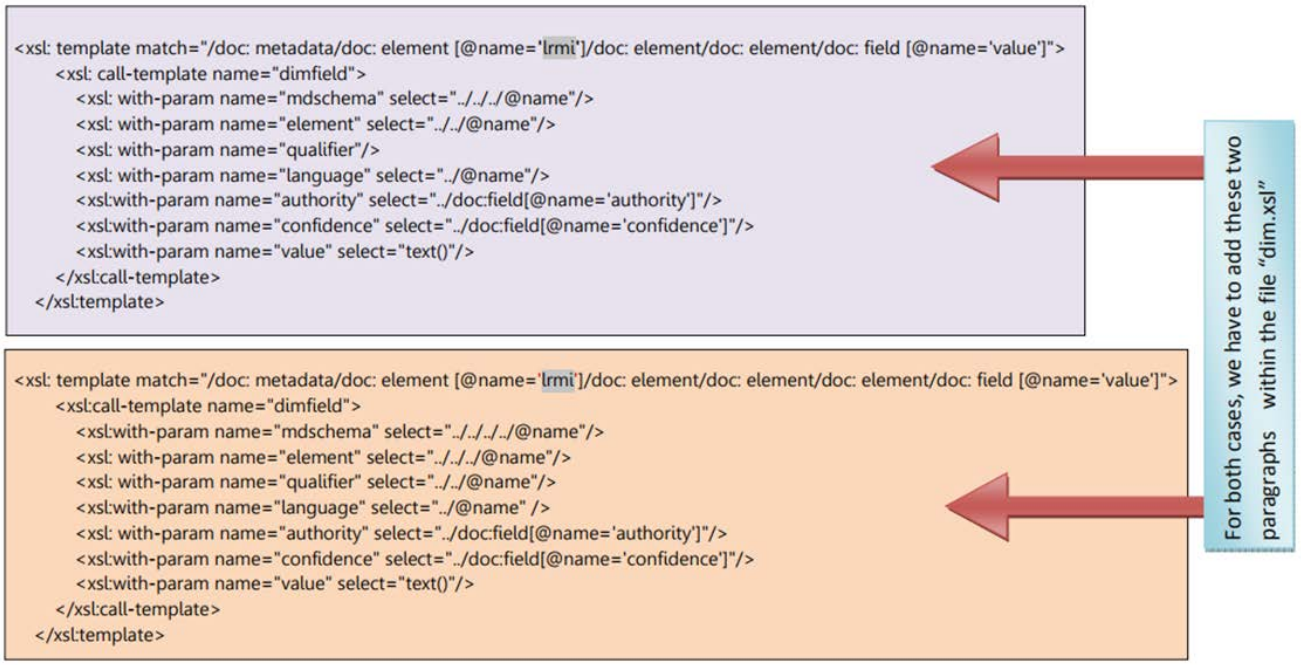

Figure 8. dim.xsl Modification.

DSpace. Custom "input-forms.xml" has been created to input "LRMI" metadata fields in DSpace.

5.4.3 Tags in 'dim.xsl'

dim.xsl

For clear understanding of the notation, go to DSpace installation directory. It could be any directory like:

'home/dspace/IDR/dspace_inst/config/crosswalks/oai/metadataFormats/dim.xsl', '/home/dspace/dspace', '/home/dspace','/usr/local/dspace' , '/dspace' or '/opt/dspace'. For example, if you come across $DSPACE_HOME/config or /home/dspace/IDR/dspace_inst/config, it indicates 'config' directory under dspace home directory where Dspace is installed.

DSpace administrator of data provider/data source server must modify the dim.xsl file for exporting additional metadata schema fields or modified fields through OAI in XML format. The path will be the following:

'home/dspace/IDR/dspace_inst/config/crosswalks/oai/metadataFormats/dim.xsl'.

The file is to be modified as below:

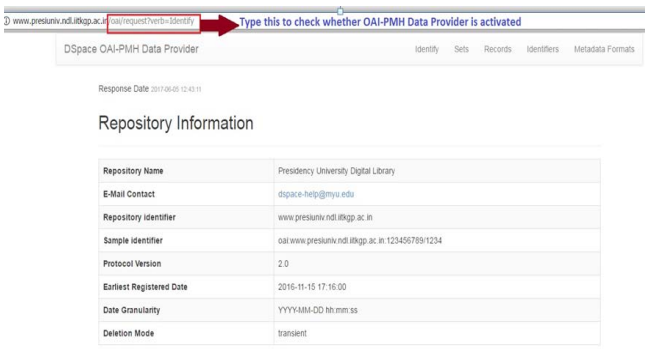

Figure 9. Exporting Additional Metadata.

5.4.4 Tags in 'oai_dc.xsl'

oai_dc.xsl

DSpace administrator of data provider/data source server must modify the oai_dc.xsl file for exporting additional metadata schema fields or modified fields through OAI in XML format. The path will be:

'home/dspace/IDR/dspace_inst/config/crosswalks/ oai/metadataFormats/ oai_dc.xsl.

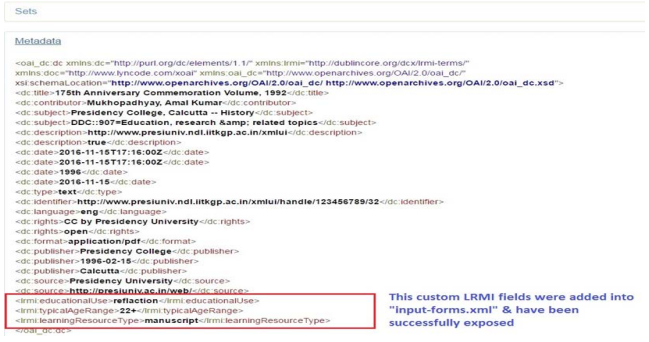

Figure 10. Modifying XSL file.

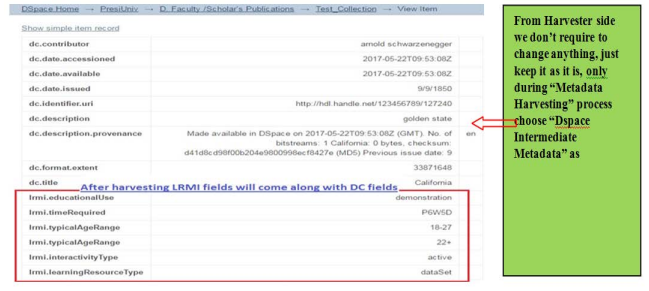

Figure 11. Extended Metadata with Additional Fields.

6. Conclusion

21st Century has seen the emergence of metadata schema for various educational objects. There are about 2-3 dozen metadata schema. As such it is difficult to ensure interoperability between all digital repositories. DSpace by default has a capability to Import/Export only "dc" fields. In this paper we explained a successful attempt implement LRMI metadata schema by modifying three files: "Input-forms.xml" 2. "dim.xsl" & 3. "oai_dc.xsl. We have harvested "lrmi" metadata fields through OAI-PMH from Data Provider Repository (DPR). Through this way an administrator can Import metadata from separate DPRs also Export metadata to separate Service Provider Repository (SPR) by OAI-PMH request in an xml format.

7. References

1. Prasad A. (2006). Implementing LOM Schema in DSpace. DRTC – ICT Conference on Digital Learning Environment. Accessed on: January 11, 2006. http://drtc.isibang.ac.in/ldl/handle/1849/221.
2. Open Archives Initiative Protocol for Metadata Harvesting. https://www.openarchives.org/pmh/
3. LRMI Metadata Terms (RDF) - Dublin Core® Metadata Initiative. http://dublincore.org/dcx/lrmi-terms/
4. IEEE SA - 1484.12.1-2002 - IEEE Standard for Learning Object Metadata. https://standards.ieee.org/findstds/standard/1484.12.1-2002.html.
5. Use the OAI-PMH interface. https://wiki.duraspace.org/display/DSPACE/Use+the+OAI-PMH+interface.
6. Prasad ARD. (2006). Interoperability and the OAI-PMH. Proceedings of the National Conference on Information Management in Digital Libraries (NCIMDIL) 2-4 Aug, 2006.
7. Deka D. (2007). OAI-PMH: A Tool for Metadata Harvesting and Federated Search. 5th Convention, Planner-2007. Gauhati University, Guwahati, December 7-8, 2007; p. 358–66. http://ir.inflibnet.ac.in:8080/ir/bitstream/1944/1360/1/40.pdf.
8. Awasthi S and Jaiswal B. (2008). Open Archives Metadata Harvesting: An Overview. 6th Convention PLANNER - 2008, 174-180.