Adding OAI-ORE Support to Repository Platforms

Alexey Maslov, Adam Mikeal, Scott Phillips, John Leggett, Mark McFarland[1]

{alexey, adam, scott, leggett}@library.tamu.edu mcfarland@austin.utexas.edu

Abstract

The Texas Digital Library is a cooperative initiative of Texas universities. One of TDL's core services is a federated collection of ETDs from its member schools. As this collection grew, the need for tools to manage the content exchange from the local to the federated repository became evident. This paper presents our experiences adding harvesting support to the DSpace repository platform using the ORE and PMH protocols from the Open Archives Initiative. We describe our use case for a statewide ETD repository and the mapping of the OAI-ORE data model to the DSpace architecture. We discuss our implementation that adds both dissemination and harvesting functionality to the repository. We conclude by discussing the architectural flexibility added to the TDL repository through this project.

1. INTRODUCTION

Data exchange between repositories is a critical component for cooperative repository initiatives. Without an automatic mechanism for content interchange, state and national repository federations encounter problems of scalability. Interoperability that includes content transmission requires established standards for describing the structure of that content.

The Open Archive Initiative's Object Reuse and Exchange (OAI-ORE) [6] defines standards and recommendations for describing and exchanging sets of digital resources. Used in combination with a discovery protocol such as OAI's Protocol for Metadata Harvesting (OAI-PMH) [2], a repository's content can be replicated at a remote location in a fully automatic manner.

This paper presents our experiences adding OAI-ORE support to the DSpace [12] repository platform. We present our use case for the Texas Digital Library (TDL): a statewide federated ETD repository. We examine the mapping between the OAI-ORE data model and the DSpace architecture. We discuss our implementation, adding both dissemination and harvesting support to the DSpace repository. Finally, we discuss future plans for this project and its contribution to the open repository community.

2. USE CASE

The Texas Digital Library (TDL) is a consortium of public and private institutions from across the state of Texas [3]. One of its earliest projects was the establishment of a federated collection of electronic theses and dissertations (ETDs) from member institutions.

Several of TDL's initial members had established repository projects using the DSpace platform. Leveraging this experience, the federated repository was built on DSpace, and utilized the Manakin interface to provide visualization and institutional branding [9][10]. However, the process of maintaining the federated collection was tedious: every semester a new batch of ETDs from each institution was uploaded to TDL using a script-assisted manual process. Changes and corrections to existing ETDs had to be replicated by hand. Within a year, this process was demonstrated to be inflexible and unable to scale with available resources as more schools joined the system.

3. SOLUTION

An optimal solution for these issues in a federated repository system contains the following properties:

- The exchange process should be programmatic and not require manual intervention by systems staff. Ideally, the federated collection would keep itself updated at set intervals.
- The process should have the ability to distinguish existing content from new items, since dropping and re-importing entire collections is infeasible as the collections grow.
- The process should support changes, corrections and the withdrawal of existing content, in addition to adding new content.
- The process must provide for the exchange of both metadata and objects.

The OAI Protocol for Metadata Harvesting has many of the features listed above. An OAI-PMH provider can be queried using a variety of parameters, allowing for selective harvesting of its content. A harvester based on the OAI-PMH protocol can restrict a search by date, allowing for retrieval of only new and updated content, as well as specific sets and metadata formats [13].

The DSpace repository platform already implements the OAI-PMH protocol to disseminate its content. Adding functionality that allows DSpace to harvest content using OAI-PMH yields a complete solution: two DSpace repositories can exchange content through an automatic and flexible mechanism. However, OAI-PMH was created specifically as an interchange method for metadata; transmitting digital objects is not part of its specification. Exchange at this level requires an additional mechanism.

3.1 Using METS

The Metadata Encoding and Transmission Standard (METS) published by the Library of Congress [4] allows for a complete representation of a digital object, including both metadata and structured references to files. DSpace includes both export and ingest hooks for METS-encoded objects; this made METS a potential solution for object interchange.

However, METS is primarily a packaging format and not an object exchange protocol. The specifics of object representation frequently vary across implementation, causing METS to lack the specificity necessary to force consistent interpretation. In the context of repository interoperability this results in a brittle solution [14].

3.2 Using OAI-ORE

Whereas METS packages metadata and object references together, the alternative is to package the object references using a specialized metadata format. The Object Reuse and Exchange protocol is a good example of this approach. OAI-ORE describes sets of Web resources in a standardized, concise manner [6]. According to ORE, a resource includes any object identified by a URI and accessible through HTTP.

In ORE terms, a DSpace repository is a set of resources. An individual DSpace item is simply a logical subset of these resources (PDF files, images, etc.) made accessible by the repository system and associated with metadata. ORE is specifically designed to describe the locations of and relationships between those resources. The necessary descriptive metadata can then be obtained separately in a different format. Following this strategy for both content and descriptive metadata, we get the minimum amount of information necessary describe everything we need.

4. IMPLEMENTATION

For the aforementioned reasons, we chose ORE to augment the OAI-PMH-based content dissemination in DSpace. To complete the circle we added functionality to DSpace to implement a fully automatic content harvester. A necessary prerequisite for implementation of either component was a mapping between the ORE data model and the DSpace architecture.

4.1 Mapping between OAI-ORE and DSpace

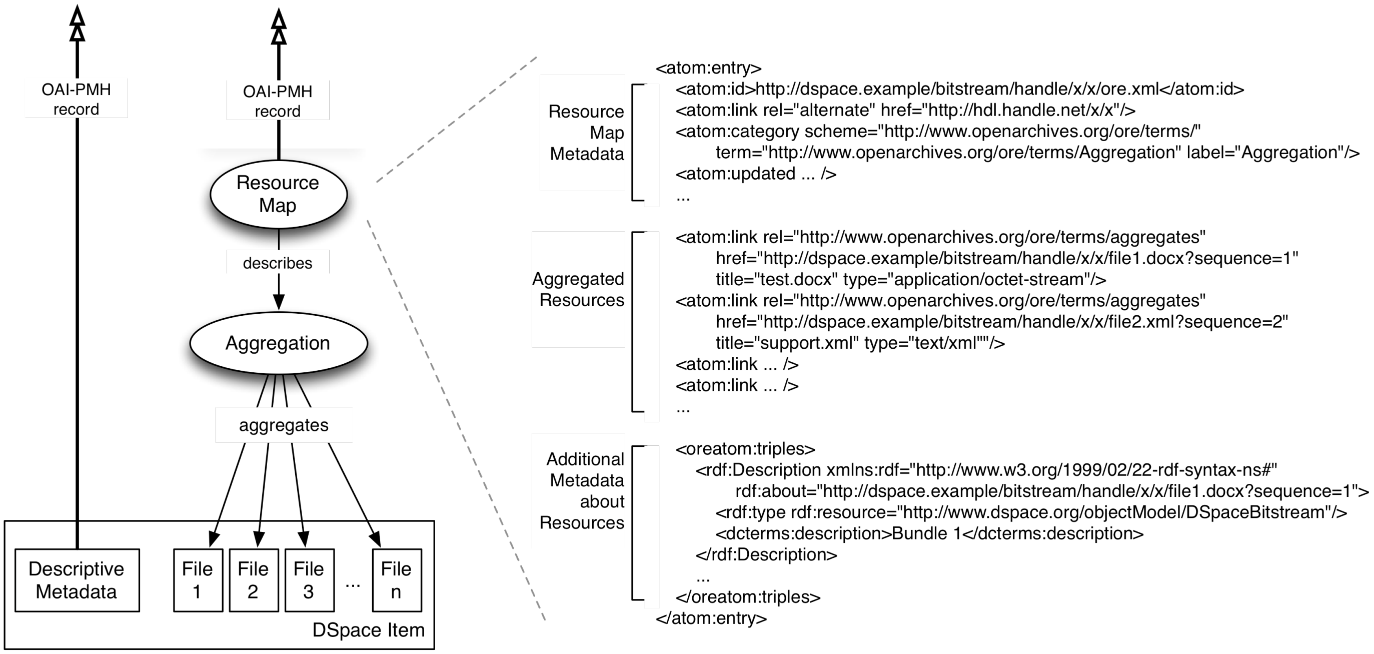
An effective mapping between OAI-ORE and DSpace means translating a DSpace Item into an ORE Aggregation and back.

As mentioned in section 3.2, the primary purpose of ORE is to describe sets of resources. The term introduced in the ORE standard for such a set is *Aggregation*, and the resources it describes are called *Aggregated Resources* [7]. In order to represent hierarchical structures, Aggregations themselves can be resources contained in other Aggregations. Since an Aggregation is an abstract concept, the ORE protocol uses a *Resource Map* to provide a concrete representation. The Atom syndication format [1], RDF/XML and RDFa [11] are the suggested serialization formats for ORE Aggregations [6].

In the DSpace architecture, an *item* is a grouping of files and descriptive metadata. The files are called *bitstreams* and are combined into abstract sets called *bundles*. There is always a primary bundle called "Content", and there may be others that store supporting files or derivative content such as thumbnails and license text. Items are grouped into larger sets called *collections* (analogous to OAI-PMH sets), which are then further grouped into nestable containers called *communities*.

The mapping between the DSpace architecture and the ORE data model for items is shown in Figure 1. Each DSpace item is an ORE Aggregation; its component bitstreams are Aggregated Resources. Moving up the DSpace hierarchy, each collection is an aggregation of items, and each community is an aggregation of collections. A Resource Map, encoded in Atom XML format, describes these Aggregations. The result is one Resource Map for each DSpace item, collection or community. Any descriptive metadata is encoded outside the ORE model as described below in section 4.2.

We selected the DSpace item as the lowest level of aggregation. An alternative model would view bundles as aggregations of bitstreams, making an item an aggregation of bundles. This model was rejected because DSpace bundles are semantically closer to metadata than sets. DSpace neither intends nor allows bundles to be used for structural organization of content; by default all



```
<atom:entry>
  <atom:id>http://dspace.example/bitstream/handle/x/x/ore.xml</atom:id>
  <atom:link rel="alternate" href="http://hdl.handle.net/x/x"/>
  <atom:category scheme="http://www.openarchives.org/ore/terms/"
        term="http://www.openarchives.org/ore/terms/Aggregation" label="Aggregation"/>
  <atom:updated ... />
  ...

  <atom:link rel="http://www.openarchives.org/ore/terms/aggregates"
        href="http://dspace.example/bitstream/handle/x/x/file1.docx?sequence=1"
        title="test.docx" type="application/octet-stream"/>
  <atom:link rel="http://www.openarchives.org/ore/terms/aggregates"
        href="http://dspace.example/bitstream/handle/x/x/file2.xml?sequence=2"
        title="support.xml" type="text/xml""/>
  <atom:link ... />
  <atom:link ... />
  ...

  <oreatom:triples>
    <rdf:Description xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
          rdf:about="http://dspace.example/bitstream/handle/x/x/file1.docx?sequence=1">
      <rdf:type rdf:resource="http://www.dspace.org/objectModel/DSpaceBitstream"/>
      <dcterms:description>Bundle 1</dcterms:description>
    </rdf:Description>
    ...
  </oreatom:triples>
</atom:entry>
```

Figure 1: Mapping between the DSpace architecture, ORE abstract data model and the final Resource Map serialization

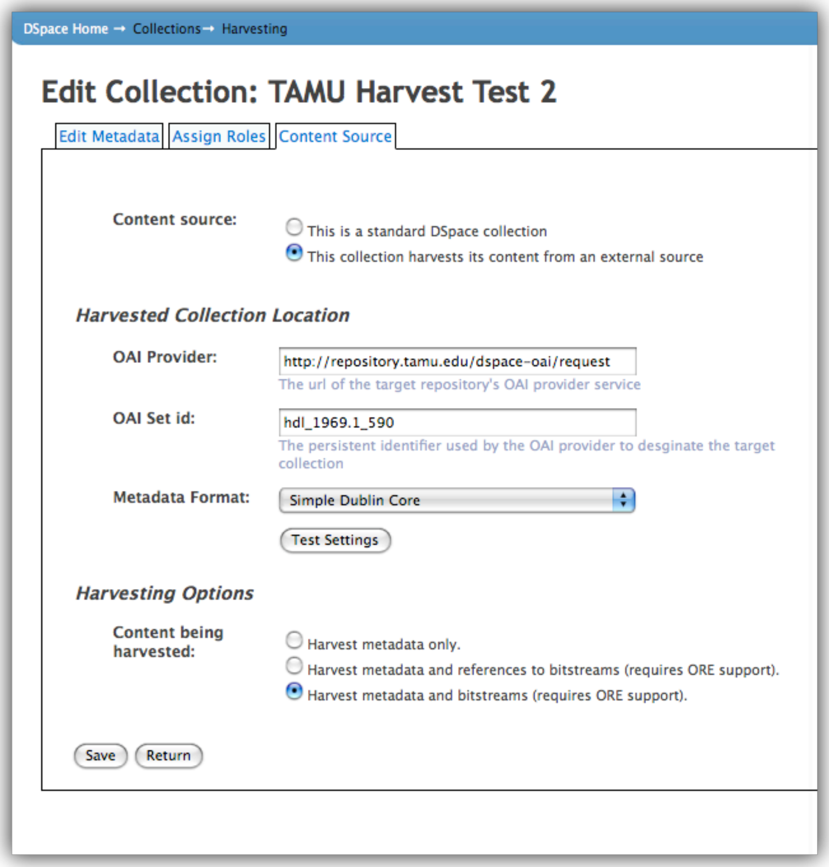
available content is stored in the same bundle. For this reason, bundle names and other details specific to DSpace are recorded in the optional metadata section of the Resource Map.

4.2 Dissemination

To disseminate content DSpace must generate resource maps for items and publish them at a persistent URI. This allows a harvester to discover and access structural information about DSpace items and their content. One of the discovery methods suggested in the ORE documentation is to embed Resource Maps inside an OAI-PMH response [8].

As mentioned in section 3, DSpace implements OAI-PMH for metadata dissemination. In this implementation, DSpace items are represented as PMH items and DSpace collections are represented as PMH sets. These items are delivered to the harvester as discrete records containing a single metadata format, such as Qualified Dublin Core, RDF or METS. To implement ORE dissemination, functionality was added to generate ORE Resource Maps and disseminate them in PMH records as another available metadata format.

Additionally, the URL space of DSpace was expanded to provide direct access to available Resource Maps independent of the PMH protocol. This allows ORE resources to maintain a persistent URI regardless of the mechanism used to generate them. All generated Resource Maps—whether contained in a PMH response or accessed directly—will still point to canonical sources.



DSpace Home → Collections → Harvesting

Edit Collection: TAMU Harvest Test 2

Content source:
○ This is a standard DSpace collection
● This collection harvests its content from an external source

Harvested Collection Location

OAI Provider: http://repository.tamu.edu/dspace-oai/request
The url of the target repository's OAI provider service

OAI Set id: hdl_1969.1_590
The persistent identifier used by the OAI provider to desginate the target collection

Metadata Format: Simple Dublin Core

Harvesting Options

Content being harvested:
○ Harvest metadata only.
○ Harvest metadata and references to bitstreams (requires ORE support).
● Harvest metadata and bitstreams (requires ORE support).

Figure 2: Collection harvesting interface in DSpace

4.3 Harvesting

DSpace already provided metadata dissemination, requiring only minor modification to extend this functionality with ORE support for object exchange. Harvesting, however, has never been part of the DSpace platform and required a complete implementation. This project consisted of three major components:

1. OAI-ORE item importer
2. OAI-PMH harvester mechanism
3. A harvest scheduling system

1. OAI-ORE item importer. DSpace needed a way to interpret ORE Resource Maps and use them to create DSpace items. This ingest component processes a Resource Map using the following algorithm: 1) it resolves the URIs to any Aggregated Resources, 2) it downloads the resources from the source location, and 3) it builds a new DSpace item, adding a new bitstream for each resource. It also scans the metadata section of the Resource Map for DSpace-specific information on bundle names. If that information is available, the bitstreams are placed in their proper bundles. Otherwise, they are placed in the default "Content" bundle.

2. OAI-PMH harvester mechanism. The item importer allows DSpace to create new items from ORE Aggregations. However, DSpace still needed a mechanism to harvest those Aggregations from remote repositories. We extended the collection management tools in DSpace to allow collection administrators to create harvested collections directly from the web interface (see Figure 2). When a collection is flagged as harvested rather than local, the administrator must provide four pieces of new information: the URL of the remote OAI-PMH provider; the set identifier of the target collection; the format to use for descriptive metadata; and whether to fetch bitstreams along with the metadata.

The harvester itself operates on the following algorithm: 1) the harvest process contacts the remote OAI-PMH provider and verifies the harvesting settings provided by the administrator, 2) it issues a PMH *ListRecords* request based on the collection's parameters and iterates over the results, and 3) for each record, the harvester a) creates a new DSpace item using the ORE item importer, b) assigns it a new local handle, c) issues a separate *GetRecord* request to obtain the descriptive metadata for that item, and d) stores a copy of the ORE resource map with the item as a hidden bitstream.

3. Harvest scheduling system. The harvest scheduler is configured on the repository level and keeps track of all harvested collections, initiating new harvest processes at set intervals. This mechanism is thread-based, and provides for several concurrent harvest processes, automating the management of the harvested collections. Once a collection is configured and verified for harvesting it becomes part of the harvesting cycle, requiring no further input from the administrator. However, options to initiate a harvest manually are still provided in all interfaces.

5. DISCUSSION AND FUTURE WORK

Adding harvesting functionality to DSpace considerably simplifies the task of maintaining TDL's federated ETD collection. However, the applications of this project extend beyond its initial use case. The ability to easily harvest content from one repository to another provides the opportunity to specialize repositories for different purposes.

For example, one repository might be a DSpace instance dedicated to the workflow of incoming ETDs using Vireo, the newly developed ETD submittal and management system [5]. The processed ETDs can then be harvested into the university's central repository. This avoids adding a layer of complexity and additional points of failure to the main repository, while still providing specialized benefits to end users.

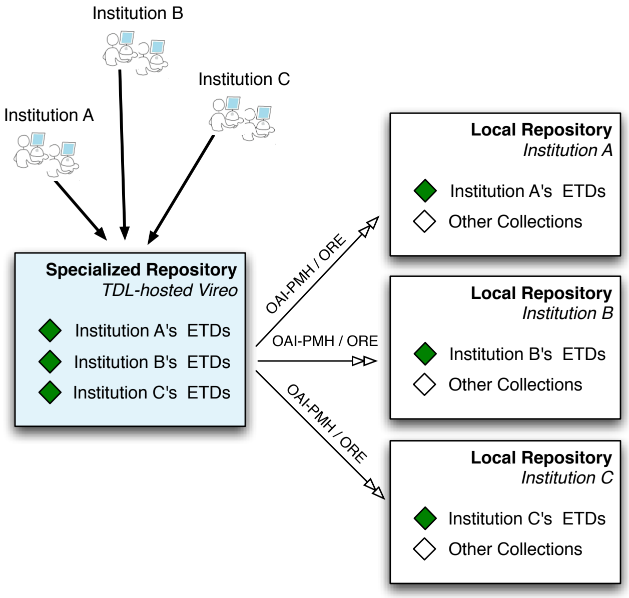


Figure 3: Example of architecture with OAI-ORE

Automatic and efficient content harvesting provides significant architectural flexibility for statewide consortium such as TDL. Smaller schools within the consortium may not wish to assume the overhead of maintaining a specialized repository instance. Automatic harvesting allows TDL to offer a hosted Vireo service and still ensure that a copy of the collection is stored locally, synchronized with the federated collection at the state level (see Figure 3). This flexibility eases the process of introducing new schools into the system.

This project will be submitted to the DSpace Foundation to be incorporated into a future release of DSpace. Having completed the initial development phase and the first round of testing within Texas, we expect to extend these tests to external data providers soon. Included in this testing is an evaluation of our harvester when accessing ORE Resource Maps generated by data providers other than DSpace.

6. REFERENCES

[1] Atom Syndication Format. <http://tools.ietf.org/html/rfc4287>

[2] Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. *The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0.* <http://www.openarchives.org/OAI/openarchivesprotocol.html>

[3] Leggett, J., McFarland, M., and Racine, D. "The Texas Digital Library: A Business Case". Prepared for and published by the Texas Digital Library, July 2005, revised July 2006.

[4] Metadata Encoding and Transmission Standard. <http://www.loc.gov/standards/mets/>

[5] Mikeal, A., Phillips, S., Koenig, J., Leggett, J., Paz, J., Brace, T., and McFarland, M. "ETD Management in the Texas Digital Library: Lessons learned from a demonstrator". In *Proceedings of the 11th International Symposium on Electronic Theses and Dissertations* (ETD08), June 2008.

[6] Open Archives Initiatives. *ORE Specifications and User Guides* (17 October 2008). <http://www.openarchives.org/ore/1.0/toc>

[7] Open Archives Initiatives. *ORE User Guide – Abstract Data Model* (17 October 2008). <http://www.openarchives.org/ore/1.0/datamodel>

[8] Open Archives Initiatives. *ORE User Guide - Resource Map Discovery* (17 October 2008). <http://www.openarchives.org/ore/1.0/discovery>

[9] Phillips, S., Green, C., Maslov, A., Mikeal, A., and Leggett, J. "Introducing Manakin: Overview and Architecture". In *Proceedings of the 2nd International Conference on Open Repositories*. January 23—26, 2007. San Antonio, TX, USA.

[10] Phillips, S., Green, C., Maslov, A., Mikeal, A., and Leggett, J. "Manakin: A New Face for DSpace". *D-Lib Magazine*, Vol. 13 No. 11, November 2007.

[11] Resource Description Framework. <http://www.w3.org/RDF/>

[12] Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., and Smith, M. "The DSpace institutional digital repository system: current functionality". In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, Houston, Texas, May 27 - 31, 2003, p87-97.

[13] Van de Sompel, Herbert and Carl Lagoze. 2002. "Notes from the Interoperability Front: A Progress Report on the Open Archives Initiative," *Lecture Notes In Computer Science. Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*. pp 144-157.

[14] Van de Sompel, H., Nelson, M.L., Lagoze, C., and Warner, S., Resource Harvesting within the OAI-PMH Framework*,* *D-Lib Magazine*, December 2004, 10(12).