Exposing OAI-PMH Metadata as Linked Data

Bernhard Haslhofer
University of Vienna
Dept. of Distributed and Multimedia Systems
Vienna, Austria
bernhard.haslhofer@univie.ac.at

Bernhard Schandl
University of Vienna
Dept. of Distributed and Multimedia Systems
Vienna, Austria
bernhard.schandl@univie.ac.at

ABSTRACT

Many institutions grant access to their metadata repositories via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). However, this protocol has two significant drawbacks: it does not make its resources accessible via dereferenceable URIs, and it provides only restricted means of selective access to metadata. The OAI2LOD Server handles these shortcomings by republishing metadata originating from an OAI-PMH endpoint according to the principles of Linked Data. As the ongoing OAI-ORE specification process shows, these principles are gaining growing importance also in the digital libraries domain.

1. INTRODUCTION

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [6] is utilised for the exchange and sharing of metadata for digital and non-digital items and enjoys growing popularity in the domain of digital libraries and archives. Currently we know of more than 1700 OAI-PMH compliant repositories exposing metadata descriptions for several millions items.

The design of OAI-PMH is based on the Web Architecture [5], but it does not treat its conceptual entities as dereferenceable resources. Also selective access to metadata is still out of its scope. One can, for instance, retrieve metadata for a certain digital item, but cannot retrieve all digital items that have been created by a certain author.

With the OAI2LOD Server we provide a possible solution for these shortcomings by following the Linked Data design principles [1] and by providing SPARQL access to metadata. The ongoing Object Reuse and Exchange (OAI-ORE) [7] standardisation indicates that the idea of Linked Data will play a substantial role in the context of digital libraries and archives. Thereby, our OAI2LOD Server could serve as bridging component between the worlds of OAI-PMH and Linked Data.

2. WHAT IS OAI-PMH?

Client applications can use the OAI-PMH protocol to harvest metadata from Data Providers using open standards such as URI, HTTP, and XML. Institutions taking the role of data providers can easily expose their metadata via OAI-PMH by implementing light-weight wrapper components on top of their existing metadata repositories.

2.1 Technical Details

The main conceptual entities in the OAI-PMH specification are Item, Record, and MetadataFormat. An item represents a digital or non-digital resource and is uniquely identified by a URI. It can be described by an arbitrary number of metadata records, each of which is bound to a certain metadata format, which can freely be chosen by the data provider. To guarantee a basic level of interoperability, all data providers must support the unqualified Dublin Core [4] format. Further, OAI-PMH provides the concept of a Set for grouping related items and their associated metadata.

OAI-PMH is implemented on top of HTTP and defines a set of verbs to request different information types: an Identify request retrieves administrative metadata (e.g., name, owner) about a repository as a whole. GetRecord is used to fetch an individual record for a certain item in a given format, whereas the request ListRecords harvests all metadata for all available items in a certain metadata format. ListIdentifiers returns the identifiers (URIs) of all available items, ListMetadataFormats the formats in which the data provider exposes metadata, and ListSets returns the available sets in an OAI-PMH repository.

Figure 1 shows a sample GetRecord request for a Dublin Core metadata record available in the Library of Congress and the corresponding response. The request URI contains the address of the repository, the verbs, and required parameters like the item URI. The response consists of a <header> section, which contains the item’s URI, and a <metadata> section encapsulating the metadata record.

2.2 Spreading and Future of OAI-PMH

There exist a number of OAI Data Provider Registries12, from which we know that currently 1765 institutions worldwide maintain OAI-PMH repositories. Regarding their application domain, we can observe that the protocol has been implemented in a variety of institutions, ranging from small research facilities to national libraries that have integrated this protocol with their catalogue systems. Examples are the Institute of Biology of the Southern Seas, exposing 403 records, and the U.S. National Library of Medicine’s digital archive, exposing 1,272,585 records.

In order to estimate the amount and the characteristics of metadata one can retrieve via OAI-PMH, we have carried out an analysis on the 915 registered repositories that delivered valid responses. Figure 2 illustrates the size of these repositories using a logarithmic scale on the Y-axis.

Copyright is held by the author(s).

LODWS April 22, 2008, Beijing, China.

1http://www.openarchives.org/Register/BrowseSites
2http://gita.grainger.uiuc.edu/registry/
REQUEST:
http://memory.loc.gov/cgi-bin/oai2_07
verb=GetRecords
identifier=oai:lcoa1.loc.gov:loc.gdc/gcfr.0018_0163
metadataPrefix=oai_dc
RESPONSE:

Figure 1: Sample OAI-PMH communication.

The results show that 843 or 92% of all repositories expose metadata for less than 20,000 items. With 14,303 being the average number of items, the total number of 13,087,842 items is made up of a large number of smaller OAI-PMH repositories.

In total, the analysed repositories expose 161 different metadata formats. Besides unqualified Dublin Core, which is required to be implemented by definition, RFC1807 (12%), MARC (11.8%) and MARC-21 (10.3%), MODS (7.5%), and METS (5.7%) are most frequently used3. The large gap between Dublin Core and the other metadata formats reveals that most data providers do not follow the OAI-PMH standard’s suggestion of exposing metadata in a semantically richer format rather than unqualified Dublin Core.

We expect the number of institutions that expose metadata via OAI-PMH to grow even further. Major attempts of building union catalogues, e.g., the The European Library (TEL) project4, rely on this protocol for indexing metadata originating from remote sources. Currently, that initiative integrates 47 national libraries and gives access to approximately 150 millions of metadata records. Since the OAI-PMH endpoints of these libraries are currently not listed in the before mentioned OAI Data Providers Registry we could not consider them in our analysis.

Another reason why we expect the number of OAI-PMH endpoints to grow is that popular open source digital library systems, such as Fedora5, DSpace6, and EPrints7, provide an OAI-PMH endpoint by default. These systems currently find a widespread adoption in various small and medium institutions (e.g., universities or museums) and will foster the global distribution of open and Web accessible metadata even more.

2.3 Shortcomings of OAI-PMH

The OAI-PMH protocol has been designed for transferring large amounts of metadata from a server to a client over the Web. From that perspective, it provides a reasonable solution for clients that need to aggregate or index metadata. However, it has two significant drawbacks:

- **Non-dereferencable identities**: although OAI-PMH is built on the Web infrastructure, we believe that it does not yet make use of its full potential. To retrieve information from a repository, a client must execute an HTTP GET request on an OAI-PMH specific URI (see Figure 1). This prevents Web clients that are unaware of the protocol specifics from accessing the repository.

- **Restricted selective access to metadata**: the record selection criteria in the OAI-PMH harvesting process are restricted to item identifiers, metadata formats, sets, and record creation date intervals. However, some clients might only be interested in records matching certain criteria (e.g., “all records describing items created by X”) or even just a subset of the available metadata values (e.g., “all authors of all books in a library”).

One could argue that these features are out of the scope of OAI-PMH and already implemented by other digital library protocols such as Z.39.598 or SRU9. However, because of the popularity and widespread adoption of OAI-PMH in contrast to other protocols, we believe that it should be enhanced in order to solve the above mentioned drawbacks.

3Further information about these standards: http://www.loc.gov/standards and http://rfc.net/rfc1807.html
4http://www.theeuropeanlibrary.org
5http://www.fedora.info
6http://www.dspace.org
7http://www.eprints.org
9http://www.loc.gov/standards/sru/specs/
Institutions, which employ the OAI-PMH, could then provide powerful metadata access functionality by implementing just a single protocol.

3. THE OAI2LOD SERVER

At a first glance, the OAI2LOD server is a wrapper that exposes metadata of OAI-PMH compliant data sources as Linked Data on the Web and provides a SPARQL query interface to these metadata. During design time we have noticed that it also covers large parts of the OAI-PMH features by simply following the Linked Data rules [1] and provides solutions for the shortcomings mentioned in the previous section.

3.1 Exposing OAI-PMH Metadata as Linked Data

The first Linked Data rule says that things should have URLs. In the context of OAI-PMH, items and sets are such things. By definition, items already fulfill that rule because, according to the OAI-PMH specification, each item must be identified by a URI (e.g., oai:lcoa1.loc.gov:loc.gdc/gcfr.0018_0163). This not the case for sets as they are identified by arbitrary strings consisting of any valid URI unrelated characters (e.g. ascfrbib). However, such strings are not valid URIs.

According to the second rule, URIs that identify resources should be resolvable HTTP URIs. In OAI-PMH it is common to use non-resolvable URNs to identify items. The OAI2LOD server bridges this gap by wrapping item URNs and set identifiers with resolvable HTTP URLs. Continuing the above example, the item’s URI becomes http://example.com/resources/item/oai:lcoa1.loc.gov:loc.gdc/gcfr.0018_0163, and the the set’s identifier becomes http://example.com/resources/set/ascfrbib.

The third Linked Data rule proposes to deliver useful information whenever a URI is dereferenced. The OAI-PMH protocol delivers useful information for harvesting clients that can parse and process OAI-PMH responses. We believe that this information might also be valuable for other human and non-human Web agents. For humans we should provide the possibility to browse, display, and search metadata using an ordinary Web browser. Other (non-human) Web agents such as Web crawlers should be able to access OAI-PMH metadata without knowing the protocol details. We fulfill this requirement (i) by assuring that the responses delivered to a client contain only resolvable HTTP URLs, and (ii) by exposing data in various representations.

When delivering metadata records to the client, we must assure that each field (e.g., creator) within a record has been assigned a resolvable URI. For some formats (e.g., Dublin Core) this is the case by definition (e.g., http://purl.org/dc/elements/1.1/creator), for others we must publish a machine-readable representation (e.g., in RDF/S or OWL) on the Web. Further, we have defined a machine-processable vocabulary defining OAI-PMH specific concepts such as Item and Set.

XHTML and RDF serialisation formats, i.e. RDF/XML and N3, are the data representations the OAI2LOD Server currently supports. While Web browsers can process the former and display the returned information to humans, the latter can be processed by machines. The server uses content negotiation, as explained in [2], to decide which representation to deliver.

In the context of OAI-PMH, the forth Linked Data rule recommends that metadata records should contain links to other related resources. One kind of link that should be included in a record delivered to a client is a reference to its origin, i.e., the OAI-PMH endpoint and all relevant protocol parameters required to retrieve the corresponding XML representation of an item and its records. We express this information using the OAI2LOD specific oai2lod:origin property, which is defined as a sub-property of rdfs:seeAlso.

Searching other OAI2LOD Server instances for equivalent or similar metadata records, is another strategy for adding links. If we refer to the example presented in Figure 1, it is quite likely that other institutions also have a copy of this book. This fact can be captured by adding an owl:sameAs property to the metadata record. Currently we do this by regarding metadata records originating from distinct server instances and comparing the values of a set of manually selected attributes according to their lexical similarity using the Levenshtein string distance [8]. If the similarity of two entries is above a certain threshold, two records are linked. In the current implementation we ask the server administrator to specify (i) target OAI2LOD Servers for linking, (ii) pairs of source and target fields to be analysed, and (iii) a similarity threshold for each pair.

Figure 3 shows the RDF/XML representation of our example metadata record as it is returned by the OAI2LOD Server. It contains the same metadata as the record in Figure 1 but represents them according to the Linked Data principles. We can see that by following the Linked Data rules, we have bridged the problem of non-dereferencable identities and support access to metadata repositories for a variety of Web agents. The other shortcoming is solved by SPARQL endpoint which allows selective record retrieval from the data stored in the OAI2LOD server.

![Figure 3: Sample OAI2LOD Server response.](http://www.mediaspaces.info/vocab/oai-pmh.rdf)
3.2 Design and Implementation

The OAI2LOD Server, as illustrated in Figure 4, is a stand-alone server implemented in Java and based on the architecture of the D2RQ Server [3]. It can be configured to expose all metadata records from a specific OAI-PMH endpoint in a certain metadata format according to the principles described above. A scheduled process regularly harvests metadata from the given endpoint, transforms them into RDF/XML using a format-specific XSL style-sheet, stores the transformed metadata in a built-in triple store, and exposes the metadata to various kinds of clients. The built-in Request Handler/Dispatcher analyses the Accept property in the HTTP headers and delivers metadata either in RDF/XML (Accept: application/rdf+xml) or in XHTML (Accept: application/xhtml+xml). It directs client requests to the OAI2LOD Server’s entry point that provides metadata in the appropriate representation using the HTTP 303 See Other response.

![OAI2LOD Request](http://www.mediaspaces.info:3030/)

![OAI-PMH Request](http://www.mediaspaces.info:3030/)

Figure 5: Comparison of OAI2LOD and corresponding OAI-PMH requests.

3.3 Preliminary Experiences

The OAI2LOD Server version 0.1 serves records from an in-memory Jena RDF model, which is fed with metadata records exposed by a certain OAI-PMH endpoint. The number of records a server instance can host, depends on the amount of memory assigned to the Java Virtual Machine.

In our test environment\(^{11}\) we have exposed 25,000 records in a JVM having 128 megabytes of RAM assigned. This indicates that a large fraction of existing OAI-PMH repositories (see Figure 2) could expose their metadata according to the Linked Data rules with very low resource effort.

3.4 Open Issues

Currently the OAI2LOD Server exposes metadata records only in a single pre-defined format. When setting up a server instance for a specific OAI-PMH repository, the administrator decides in which format the metadata records are harvested. Since this approach contradicts a central idea of OAI-PMH we will further investigate how the OAI2LOD Server could serve metadata in multiple formats. One potential solution is to define mappings between formats.

Another important OAI-PMH feature is batch retrieval of metadata records. Using the ListRecords request, a client can iteratively retrieve a chunk of records. The OAI2LOD Server currently supports these features through SPARQL and its LIMIT and OFFSET clauses. However, we believe that alternatively we could offer that feature via a dereferencable URI.

The OAI2LOD Server’s capabilities of linking items with other resources on the Web are limited and still rely on human intervention. We need to experiment with further duplicate detection algorithms and similarity metrics, in order to achieve better and scalable results.

4. OAI-ORE

The Open Archives Initiative Object Reuse and Exchange (OAI-ORE) [7] specification is the latest standardisation effort driven by the designers of the OAI-PMH protocol. Although the standards are still in an alpha release status, we can already notice strong similarities with the ideas of

\(^{11}\)http://www.mediaspaces.info:3030/
Linked Data and the OA2LOD Server respectively.

OAI-ORE is a set of standards for the description and exchange of aggregations of Web resources. A resource can be anything that is identified with a URI such as Web sites, online multimedia content, or items stored in institutional digital library systems. In theORE data model an aggregation is an instance of the conceptual entity Resource Map and is identified by a URI. A resource map describes the encapsulated resources as a set of machine readable RDF statements, which makes them readable for a variety of Web agents. Clients can retrieve aggregations by executing an HTTP GET request on a resource map’s URI. The ATOM Syndication Format\textsuperscript{12} is specified as the primary serialisation format for delivering resource maps to clients. However, since the ORE data model is defined in RDF, resources can not only be mapped to the ATOM format but also serialised in other RDF exchange formats such as RDF/XML or N3.

Regarding the OAI-ORE specification from the perspective of Linked Data, we can observe that the first two Linked Data rules are fundamental building blocks of the standard: all things, i.e., resource maps and the aggregated resources, are identified by dereferencable URIs. Further, all terms used for describing aggregations have a well-defined semantics, published in terms of a Web accessible vocabulary definition. It also considers the third rule because resolving the URIs returns useful—i.e., processable and interpretable—information for both human and machines. Finally, OAI-ORE also follows the fourth rule by providing several possibilities to link resources: first, an aggregation of resources is by definition a collection of linked (ore:aggregates) resources; second, the ORE model uses the owl:sameAs property to denote that two identifiers refer to the same information object; third, it supports the concepts of nested aggregations.

OAI-PMH and OAI-ORE overlap in the fact that Resource Maps can be included as metadata records in OAI-PMH responses, which allows batch retrieval and harvesting of aggregation information. We believe that there lies a great potential in a tighter integration of these two standards: if OAI-PMH metadata repositories expose their items as Web resources by assigning them HTTP-dereferencable URIs, these items could take part in OAI-ORE aggregations. One possible strategy could be to define a common core data model that links these two standards so that the ORE specification builds on top of the OAI-PMH protocol. Meanwhile, the OA2LOD Server can serve as a bridge between these two standards.

5. CONCLUSION

In this paper we have presented the OA2LOD Server, a software component that republishes metadata from OAI-PMH compliant repositories according to the Linked Data principles. It fulfills two major purposes: first it exposes the conceptual OAI-PMH entities (item, set) as dereferencable Web resources, and second, it provides selective access to metadata via a SPARQL endpoint. These features make OAI-PMH metadata accessible also for Web clients not being aware of the OAI-PMH protocol specifics.

Since the alpha version of the OAI-ORE specification has been released, we can observe that also in the digital libraries domain the Linked Data principles will play an important role. Also for the already established OAI-PMH protocol, it would make sense to treat its conceptual entities (items, sets) as resources that can be dereferenced via URIs. In that way, they could take part in OAI-ORE aggregations. Meanwhile, the OA2LOD Server can be used for bridging the conceptual gap between these standards.

Our work on the OA2LOD Server will continue: first we will deal with the open issues mentioned in Section 3.4. Second, we will investigate techniques for linking metadata and third, we also plan to implement OAI-ORE support for aggregating items.

6. REFERENCES
