A distributed registry for OpenURL metadata schemas with an OAI-PMH conformant central repository

Herbert Van de Sompel
Los Alamos National Laboratory
Los Alamos, NM, US
herbertv@lanl.gov

Donna Bergmark
Cornell University
Ithaca, NY, US
bergmark@cs.cornell.edu

Abstract

This paper describes a distributed registry of XML Schemas in which a central facility uses the Open Archives Protocol for Metadata Harvesting (OAI-PMH) as a communication mechanism between registrars and parties that need to remain up-to-date regarding registration, updating and deregistration of those Schemas. This novel application of the OAI-PMH is described in the context of registration of XML Schema for usage in OpenURL applications. It suggests the applicability of the protocol in areas that go well beyond resource discovery, and that may actually reach into the realm of synchronization of administrative networked processes, and digital preservation.

1. Introduction

The OpenURL framework is an interoperability model that facilitates context-sensitive reference linking in distributed digital libraries [1]. When delivering metadata descriptions to users of a digital library, information providers introduce an OpenURL for each such description. The OpenURL is an HTTP GET/POST request targeted at a linking server of the user's preference. When clicked by the user, the OpenURL transfers relevant parts of the metadata description to the user's linking server. This transfer is either:

- By-value - the metadata is delivered in the HTTP GET/POST, or
- By-reference - a pointer to the metadata is delivered in the HTTP GET/POST.

Upon receipt, the linking server analyzes the metadata and delivers appropriate linking services to the user.

Whereas the draft OpenURL specification [2] was focused on facilitating the on-demand transfer of metadata of scholarly works, the emerging OpenURL NISO standard aims to be applicable well beyond this realm. To that end, an extensibility mechanism is being built in that allows by-value as well as by-reference transfer of metadata expressed according to multiple metadata formats, subject to the following restrictions:

- Only a single by-value and a single by-reference format can be present on any one OpenURL;
- By-value and by-reference metadata formats must be defined by means of an XML Schema [1].

As a result, linking servers will need access to these XML Schemas. As a matter of fact, in order to be able to cope with the input of OpenURLs carrying metadata expressed according to a certain format format, a linking server must have access to the Schema defining that format well before the receipt of such OpenURL. Therefore, a facility that allows linking servers to remain informed regarding XML Schemas used for OpenURL is required. This paper proposes a low-overhead approach for the implementation of such a facility, which introduces a novel application of the Open Archives Protocol for Metadata Harvesting (OAI-PMH) [3].

2. The Registry Model

The core design goal of the proposed registration facility (Figure 1) is ease of administration for all parties involved. In order to achieve this goal, the model federates the administration of the registration service into:

- Distributed nodes - Communities that define XML Schemas for usage with the OpenURL carry the responsibility of maintaining those schemas in a network accessible location under their control;
- A central repository – A registration services acting as a gateway between the distributed nodes and the linking servers that need access to the Schemas.

The central node provides the following functionality to the distributed nodes:

- A facility to register XML Schema;

[1] This restriction has not yet been approved by the NISO-AX Committee, but is likely to be adopted.

Proceedings of the International Conference on Parallel Processing Workshops (ICPPW'02)
1530-2016/02 $17.00 © 2002 **IEEE**

- A validation of the XML Schema comprising XML Schema validation and validation of additional Open specific requirements [2].

The central repository provides linking servers with an OAI-PMH harvesting interface that supports:

- Identification of newly registered, updated and de-registered Schema;
- Retrieval of up-to-date mirror copies of registered Schemas.

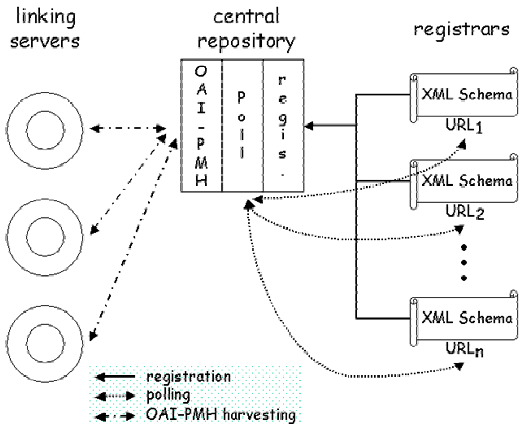

Figure 1. The registry model

3. (De) Registration of a Schema

Communities register new XML Schemas, by making them accessible via a web-server under their control. Next, the Schemas are registered with the central repository. As a result of registration, the central repository minimally holds the following information regarding a registered Schema:

- The network location of the Schema;
- The date of registration;
- A mirror copy of the registered Schema.

De-registration of a Schema is performed by removing the Schema from the registered network location of the Schema.

4. The OAI-PMH Conformant Repository

The central repository is organized as an OAI-PMH conformant repository with the following core properties:

- The *resources* it contains metadata about are the registered XML Schemas;
- The *items* that – in the OAI-PMH – are the gateway to metadata about resources carry as *identifiers* the network location of the registered Schemas;
- The repository supports *records* expressed according to the XML Schema for the mandatory unqualified Dublin Core (oai_dc), according to the XML Schema defining XML Schema (xsi), and according to a newly created format (poll) that facilitates expressing process-related information. Table 1 shows the response to the ListMetadataFormats OAI-PMH request reflecting these supported formats;
- The *set* structure of the repository reflects the OpenURL version and usability of Schema for by-value and/or by-reference transfer of metadata on the OpenURL. Table 2 shows the response to ListSets;
- The repository supports *deletions* in a *persistent* manner.

Table 1. Excerpt of ListMetadataFormats response

```
<ListMetadataFormats>
 <metadataFormat>
  <metadataPrefix>oai_dc</metadataPrefix>
   <schema>
    http://www.openarchives.org/OAI/2.0/oai_dc.xsd
   </schema>
   <metadataNamespace>
    http://www.openarchives.org/OAI/2.0/oai_dc/
   </metadataNamespace>
 </metadataFormat>
 <metadataFormat>
  <metadataPrefix>xsi</metadataPrefix>
  <schema>
   http://www.w3.org/2001/XMLSchema.xsd
  </schema>
  <metadataNamespace>
   http://www.w3.org/2001/XMLSchema-instance
  </metadataNamespace>
 </metadataFormat>
 <metadataFormat>
  <metadataPrefix>poll</metadataPrefix>
  <schema>
   http://www.openurl.info/poll.xsd
  </schema>
  <metadataNamespace>
   http://www.openurl.info/poll/
  </metadataNamespace>
 </metadataFormat>
</ListMetadataFormats>
```

As a result, for a given item, the repository can simultaneously contain:

- An oai_dc-record (Table 3): Registration information. The *datestamp* of this record reflects the date of registration of the Schema. This datestamp will never change, unless the Schema is deregistered. At that

[2] For instance, element names for by-value delivery of metadata must not contain the underscore [_] character.

moment, the datestamp will be updated and the oai_dc-record will receive a "deleted" status.

- An xsi-record (Table 4): A mirror copy of the Schema at the registered schema location. The datestamp of this record originally reflects the date of registration of the Schema. This datestamp will be updated as a result of the central repository's tracking of updates to and deregistration of registered Schema, as explained hereafter.
- A poll-record (Table 5): This record logs the results of the repository's recurrent polling of registered Schemas, as explained below.

Table 2. Excerpt of ListSets response

```
<ListSets>
 <set>
  <setSpec>openURL</setSpec>
  <setName>Metadata schema for usage
   with OpenURL</setName>
 </set>
 <set>
  <setSpec>openURL:1.0</setSpec>
  <setName>Metadata schema for usage
   with OpenURL v.1.0</setName>
 </set>
 <set>
  <setSpec>openURL:1.0:by-value</setSpec>
  <setName>by value metadata formats for
   OpenURL v.1.0</setName>
 </set>
 <set>
  <setSpec>openURL:1.0:by-ref</setSpec>
  <setName>by reference metadata formats for
   OpenURL v.1.0</setName>
 </set>
</ListSets>
```

The central repository recurrently polls the URLs of registered Schemas. The following main scenarios are possible:

- HTTP status-code 200: In case the repository is able to fetch the Schema at the registered network location, it will compare it to the mirror copy it holds. If the fetched Schema and the mirror copy are identical, no further action is undertaken. If the comparison technique reveals a difference, the fetched Schema will replace the existing mirror copy held in the repository. At the same time, the datestamp of the xsi-record will be updated. The datestamp of the oai_dc record remains unchanged. A new entry is made in the poll-record, reflecting the successful polling of the Schema. The datestamp of the poll-record is updated.
- HTTP status-code 404: In case the repository is unable to fetch the Schema, deregistration is assumed. At this point, the oai_dc-record receives a "deleted" status, and its datestamp is updated. Subject to policy decisions, the mirror copy of the Schema can be kept in the repository for archival purposes or it can be deleted. In the first case, the datestamp of the xsi-record remains unchanged. In the second case, the xsi-record receives a "deleted" status, and its datestamp is updated. A new entry is made in the poll-record, reflecting the unsuccessful polling of the Schema. The datestamp of the poll-record is updated. It is likely that the central repository would only expose deregistration of a Schema (deleted oai_dc record) after several unsuccessful polls. The information contained in the poll-record will support the repository to make decisions with this respect.

Table 3. oai_dc-record; schema registration 2002-02-04

```
<record>
 <header>
  <identfier>
   http://www.openurl.info/journal.xsd
  </identifier>
  <datestamp>2002-02-04</datestamp>
  <setSpec>openURL:1.0:by-value</setSpec>
 </header>
 <metadata>
  <oai_dc:dc>
   <dc:creator>herbertv@lanl.gov</dc:creator>
   <dc:description>Schema for journal
    metadata for use in OpenURL v.1.0
   </dc:description>
   <dc:identifier>
    http://www.openurl.info/journal.xsd
   </dc:identifier>
   <dc:date>2002-01-01</dc:date>
  </oai_dc:dc>
 </metadata>
</record>
```

5. The Linking Servers

Resulting from this organization of the central repository is the possibility for linking servers to use OAI-PMH requests to remain informed about the comings and goings of XML Schemas for usage on the OpenURL:

- The poll-records provide information regarding the central repository's recurrent polling of registered Schemas. Its datestamp reflects the time of the most recent polling-activity.
- Through their datestamps, oai_dc-records provide information about registration and deregistration of Schemas.
- The xsi-records provide up-to-date mirror copies of registered Schemas. Their datestamp reflects registration of and updates to Schemas. Depending on policy decisions, they may also reflect deregistration.

Table 4. xsi-record; schema update 2002-05-13

```
<record>
 <header>
  <identfier>
   http://www.openurl.info/journal.xsd
  </identifier>
  <datestamp>2002-05-13</datestamp>
  <setSpec>openURL:1.0:by-value</setSpec>
 </header>
 <metadata>
 <xsi:schema targetNamespace=
  "http://www.openurl.info/journal/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified"
  attributeFormDefault="unqualified">
<xsi:element name="journal" type="jo:journalType" />
<xsi:complexType name="journalType">
<xsi:choice maxOccurs="unbounded">
<xsi:element name="jtitle"
   type="jo:journalTitleType" />
<xsi:element name="stitle"
 type="jo:shortJournalTitleType" />
<xsi:element name="ISSN" type="jo:ISSNType" />
...
</metadata>
</record>
```

6. Conclusions

The registration facility described above meets the goal of keeping linking servers up-to-date on Schemas for usage with the OpenURL. It does so in a manner that makes implementation and support simple for all parties involved:

- Communities store their Schemas on a site under their control, and as a result can update or deregister them without the need to go through a central agency. This avoids the implementation of administrative services by the central registry that keeps track of editorial rights to registered Schemas. Registration boils down to entering a URL in a Web form. Deregistration is as simple as removing the Schema from that URL.
- The central repository can use off-the-shelf public domain tools to implement the polling facility and the OAI-PMH repository. Operation of the service is expected to be automatic and hence to come at low cost.
- The linking servers can use off-the-shelf, public domain OAI-PMH harvesting tools to keep track of the doings of registered XML Schemas.

The described facility suggests using the OAI-PMH in a manner that is different from its common application for resource discovery. In the model presented, it is applied to support communication between parties that are not directly connected. More specifically, it is used for capturing and exposing events occurring to a collection of distributed documents, via the intermediation of a trusted party. In that sense, the model might be applied to report preservation risks regarding web-documents to a community of digital archives [4] [5].

Table 5. poll-record: most recent poll 2002-07-22

```
<record>
 <header>
  <identfier>
   http://www.openurl.info/journal.xsd
  </identifier>
  <datestamp>2002-07-22</datestamp>
  <setSpec>openURL:1.0:by-value</setSpec>
 </header>
 <metadata>
  <poll>
   <fetch status="200">2002-07-22</fetch>
   <fetch status="200">2002-07-20</fetch>
   <fetch status="404">2002-07-18</fetch>
   <fetch status="200">2002-07-16</fetch>
   <fetch status="200">2002-07-14</fetch>
    ...
  </poll>
 </metadata>
</record>
```

7. References

[1] Van de Sompel, H. and Beit-Arie, Oren. 2001. Open Linking in the Scholarly Information Environment Using the OpenURL Framework. D-Lib Magazine. available at <http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>

[2] Van de Sompel, Herbert et al. 2000. OpenURL syntax description. available at <http://www.sfxit.com/openurl/openurl.html>

[3] Lagoze, Carl, Van de Sompel, Herbert et al. (editors) 2002. The Open Archives Protocol for Metadata Harvesting version 2.0. available at <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>

[4] Kenney, A.R. et al. 2002. Preservation Risk Management for Web Resources. D-Lib Magazine. available at <http://www.dlib.org/dlib/january02/kenney/01kenney.html>

[5] Nelson, M.L. and Allen, B.D. 2002. Object Persistence and Availability in Digital Libraries. D-Lib Magazine. available at <http://www.dlib.org/dlib/january02/nelson/01nelson.html>