elpub2002 Proceedings — J. A. Carvalho et. al. (Eds.) — VWF Berlin, 2002

Creating Open Digital Library Using XML

Implementation of OAi-PMH

Martin Vesely, Thomas Baron, Jean-Yves Le Meur, and Tibor Simko

CERN, ETT-DH Division, CH-1211 Geneva 23
Switzerland

Abstract. This article describes the implementation of the OAi-PMH protocol within the CERN Document Server (CDS). In terms of the protocol, CERN acts both as a data provider and service provider and the two core applications are described. The application of XML Schema and XSLT technology is emphasized.

1 Introduction

1.1 OAi-PMH

The protocol for metadata harvesting (PMH) has been developed within the Open Archives initiative (OAi) aiming to interconnect archived collections of electronic journal articles and preprints in different scientific disciplines. The OAi-PMH [1] offers a framework for XML-based metadata exchange among heterogeneous databases – or technologically other types of storage – referred to as metadata repositories spread all over the Internet network [2]. Applications layered on top of this protocol enable the creation of open digital library services. However, by its character, the protocol has an impact beyond the area of digital libraries and concerns all communities that are engaged in publishing content on the Web and other communities where metadata can be shared with a benefit.

The OAi metadata harvesting is based on the client-server architecture, where the harvester (client) issues OAi-requests towards data provider's metadata repository (server) that in turn sends the OAi-responses back to the harvester. The OAi-PMH protocol is located in the upper part of the application layer of the TCP/IP reference model and specifies which transfer protocol is used to facilitate the metadata transfer and focuses on how the syntactic, structural and semantic interoperability between metadata repositories is provided.

1.2 Open Digital Library

Open digital library is an electronic document handling system that has implemented an interface for the data exchange conforming to some accepted standard. This term is also used by other authors to describe extensibility of open protocols applied in the field of digital libraries [3] or to define a distributed or component-based digital library [4]. These different notions, however, have a common point that is essential to all of them: the interoperability of digital library systems.

At CERN the scientific library acts both as a data provider and a service provider at the same time. Within the OAi community such a metadata repository is often labeled an 'aggregator'. In case some value-added processing [5] takes place on top of harvested and re-exported records, we rather use the term 'brokering' evoking the terminology introduced by [6].

The CERN Document Server as a metadata broker is harvesting and indexing records from multiple repositories and maintaining them by value-added processing. Currently some 2000 preprints per week are harvested. Metadata is then exposed for further harvesting by third parties. This approach introduces new architectural issues. *Hierarchical harvesting* using OAi-PMH was first implemented in the Arc system at the Old Dominion University of Virginia, USA [7].

By *reciprocal harvesting* we denote a situation that emerges when exported records are harvested back by the originating metadata repository. Both repositories are then equivalent in the tasks they perform in terms of OAi-PMH functionality. In a digital library, metadata records are constantly being maintained and modified. An example of such modification can be a change of the record status from a preprint to a published article. A record that was harvested from another metadata repository and then updated locally can in turn be of interest to the originating metadata repository. Within this paradigm there has to be a mechanism that allows the determination of the up-to-date record. A metadata repository that modifies a record should attach a local OAi identifier to it. In order to keep track of the changes performed and to assure the identification of the record and its uniqueness within a metadata repository, the provenance information has to be kept and attached to the re-exported record.

1.3 OAi / XML

OAi-PMH is an application of XML by employing XML-related specifications such as XML Schema and XSL Transformations. Particular aspects for XML application within digital libraries are described for example in [8]. The OAi-response is represented by an XML-encoded byte stream, where the structure of OAi-responses and the internal structure of the metadata container are defined by an XML Schema. The XML document must comply to the default XML declaration, which conforms to XML version 1.0 and requires the UTF-8 encoding.

The metadata itself is wrapped in a metadata container that is carried inside the XML. Metadata can be disseminated in various metadata formats that are specified by its XML Schema. The default metadata format is based on the Dublin Core Metadata Element Set [9]; the other defined formats include RFC1807 [10] and MARC21 [11].

2 Implementation at CERN

2.1 Motivation

The CERN Document Handling and Scientific Information Service groups have been maintaining a repository of High Energy Physics (HEP) documents for more than four decades providing information service to scientists at CERN and to the worldwide HEP community. The repository contains many types of documents, mainly preprints in electronic form (eprints) forming the CERN eprint archive, published articles, books, theses, conference proceedings and presentations, large collection of multimedia material like photos and videotaped lectures [12]. The system is run by an in-house developed software – CERN Document Server Software [13] that enables highly customizable functionality for document handling, metadata management and information service delivery.

A cooperation within the eprints community led to a need of practicing the metadata exchange between the repositories on a regular basis [14]. The original metadata harvesting method employed an in-house developed modular system that offered a unified way of uploading heterogeneous semi-structured metadata [15] into the local metadata repository. The heterogeneities, however, were not solved on the level of the application protocol, but rather on the level of the application itself. Each resource had to be precisely described, including the syntactical description of the metadata structures it delivered. This requirement has inevitably led to a high configuration complexity of the tool, where an application of OAi-PMH protocol would offer a more efficient alternative.

Within the CDSware the MARC XML [16] is used as data interface between different applications that take care of gathering, storing, maintaining and retrieving of metadata from the database. Having accepted the OAi-PMH approach, the XML MARC is also used as a data interface with external applications. Apart from the harvesting approach there are also other metadata acquisition techniques employed such as the WebSubmit application that permits the authors and librarians to directly submit their metadata.

2.2 CERN as Metadata Provider

The OAi metadata repository is built as an interface to CDSware [17] bibliographic database management system implemented on top of MySQL RDBMS and Apache

web server. In the database records are stored in the MARC21 metadata format. The XML byte stream is created as an output of the database search engine. By default the metadata format used is the MARC XML directly transformed from MARC21. When another metadata format is requested the processor creates it on the fly by an integrated function or by applying the appropriate XSL stylesheet for the transformation into requested metadata format. When the repository receives an OAi-request the following tasks are performed within the CDSware:

- OAi request parsing and validation
- Flow control management
- Composition of the database query and querying the database
- Fetching of matched records
- Creation of metadata container in the requested format
- Validation of the created XML file
- Transfer of the OAi response via HTTP

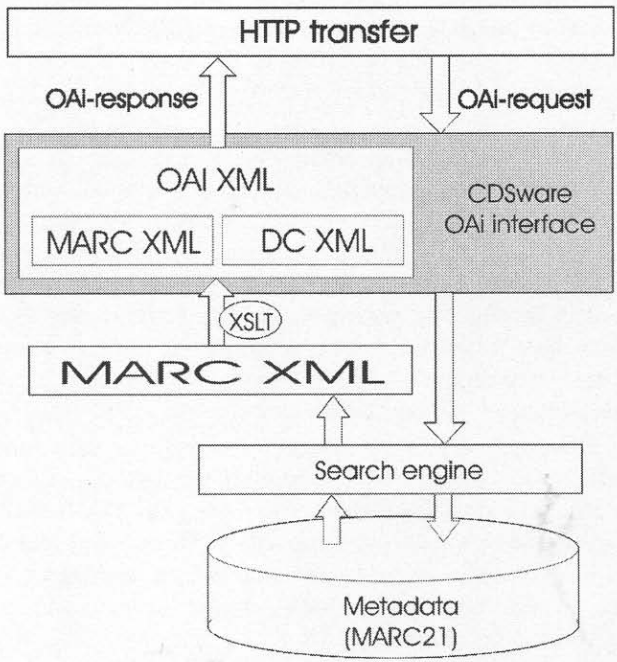

Fig. 1. *CDSware OAi-interface*

The CDSware metadata repository implementation (Figure 1) benefits from the OAi sets specification that enables it to divide the entire repository into partial

collections. There are two ways how OAi sets can be created within the CDSware metadata repository. Either they can be directly related to the defined metadata collection, or they can be defined dynamically using any search query.

The flow control is implemented using a temporary output stream caching. If the OAi-request cannot be satisfied by one OAi-response the request parameters are temporarily stored in the file system. The information includes the entire original OAi-request, the complete list of matched record identifiers belonging to the resulting set and the resumption token as a pending request identifier. This information is kept until the expiration date and time in order to be able to deal with potential transfer failures. Records are only fetched from the database as consecutive requests are being received. The consecutive requests do not invoke any additional database querying, the output stream of previously executed query is processed instead.

2.3 CERN as Metadata Harvester

CDSware metadata harvester is used to harvest raw metadata from selected remote repositories. Harvests are being performed depending on the selected harvesting mode (full, incremental).

When an OAi-request is received the following tasks are performed:

- Request parsing and composition of OAi-request
- Reception of metadata load
- Transformation of metadata format
- Pre-processing of metadata records for final upload
- Upload of records into the local database

The semantic interoperability is provided by application of the Dublin Core metadata element set namespace [18]. However, the Dublin Core standard as such does offer a limited framework by supporting only a fifteen of basic metadata elements such as creator, title, description, etc. Therefore the interoperability could preferably be achieved on the level of XSL transformations between the native formats and some standard *rich* metadata element set such as MARC21. The XML Schema for MARC21 and related transformation stylesheets have recently been standardized by the Library of Congress.

In CDSware the harvester application is represented by a modular system, implemented partially in Python and PHP. The **harvester** module provides an implementation of the transfer subsystem of the OAi-PMH. It takes care of the consistent data transfer including the flow control issues. The consistency is assured in a way that no incomplete response is considered as reliable and as such is not forwarded to further processing. All partial responses are assembled upon the final response has been received. The harvester can be invoked by a scheduled job or manually via the

user interface. The **converter** module performs the conversions between various metadata formats using XSL transformations over harvested metadata or other conversions. Apart from the OAi compliancy, the **converter** is used to process metadata from other than OAi-compliant repositories. The module was extended by a general metadata parser, which allows processing of semi-structured data input independent on XML. After this processing records are kept in the MARC XML format until they are finally loaded by **uploader**. The uploader utility creates new records or updates the existing ones in the bibliographic database with harvested and pre-processed metadata.

On the top of the actual harvesting procedure there is a need to define a harvesting strategy that states what and when should be harvested. The CDSware metadata harvester enables to define such strategies in the administration module, where the system administrator can specify a harvesting schedule for periodical incremental harvests as well as for individual full harvests. This administration module provides a way to keep the metadata repositories up-to-date and synchronized.

3 Conclusions

We have shown in this paper how the XML technology is successfully applied within the OAi-PMH framework and how XML and XSLT techniques are used within the CERN Document Server. The OAI-PMH emerges today as the best and standard solution for exchanging, merging and adding value between heterogeneous metadata repositories and online digital libraries. We believe that a distributed institution-based metadata repositories, interacting via the OAi-PMH protocol, will enhance and ease the comprehensive coverage of scientific literature in many areas.

Currently, 90% of eprints metadata entered in CERN Document Server is received using the harvesting approach, representing about 100 000 documents per year. In 2001, 70% came from the "central" self-archiving system (arXiv.org) and the rest either from CERN (via direct submission) or from about 80 information sources at 30 different HEP institutes. The ideal state would be achieved if the OAi-PMH were implemented on top of all institution databases. XML MARC standard is recommended to prevent information loss involved in using the "least common denominator" standards such as XML Dublin Core.

XML applications described in this paper are freely available as part of the CERN Document Server Software suite.

References

1. Open Archives Initiative Release Version 2.0 of the Protocol for Metadata Harvesting, http://www.openarchives.org/news/oaiv2press020614.html
2. Lagoze C, van de Sompel H: The Open Archives Initiative: Building a low barrier interoperability framework, Proceedings of the ACM/IEEE Joint Conference on Digtial Libraries, Roanoke VA, June 24-28, 2001, pp. 54-62
3. Suleman H, Fox E A: A Framework for building Open Digital Libraries, D-Lib Magazine 12/2001 http://www.dlib.org/dlib/december01/suleman/ 12suleman.html
4. Kahn R, WilenskyR: A Framework for Distributed Digital Object Services, Coporation for National Research Initiatives, Reston, Working Paper cnri.dlib/tn95-01, 1995. http://www.cnri.reston.va.us/home/cstr/arch/k-w.html
5. Kashyap V, Sheth A: Information Brokering Across Heterogeneous Digital Data: A metadata-based Aproach, Kluwer Academic Publishers http://lsdis.cs.uga.edu/lib/download/kluwerbook.pdf
6. Bowman C M, Danzig P B, Hardy D R, Manber U, Schwartz M F: The Harvest Information Discovery and Access System, Computer Networks and ISDN Systems, 28/1995
7. Zubair M, Maly K, Liu X: Arc, OAI Service Provider for Digital Library Federation, D-Lib Magazine, April 2001, vol. 7, N. 4, http://www.dlib.org/dlib/april01/liu/04liu.html
8. van Hervijnen E: The Impact of XML on library procedures and services, HEP Library Webzine, 1/2000, http://library.cern.ch/HEPLW/1/papers/2/
9. Dublin Core Metadata Element Set, v1.1, Reference Description, http://dublincore.org/documents/1999/07/02/dces/
10. RFC 1807, http://www.ietf.org/rfc/rfc1807.txt
11. MARC21, http://www.loc.gov/marc
12. CERN Document Server, http://cds.cern.ch/
13. CERN Document Server Software, http://cdsware.cern.ch/
14. Pignard N, Geretschlager I, Jerdelet J: Le traitment informatise de ressources electroniques au CERN, Documentaliste Science de l'information, n. 1, vol. 38, 2001
15. Vesely M: Using Internet/Intranet Technologies in Library Automation, Thesis, Prague University of Economics, 2000
16. Library of Congress, http://www.loc.gov/
17. Vesely M: CERN Document Server Software, Invited talk, 1^{st} OAF Workshop, 13-14 May 2002, Pisa, Italy
18. Namespace Policy for the Dublin Core Metadata Initiative (DCMI), http://dublincore.org/documents/2001/10/26/dcmi-namespace/