

OPEN ARCHIVES FORUM:

EVALUATIVE REVIEW OF METADATA HARVESTING PILOT

Project Number:	IST-2001-32015
Project Title:	Open Archives Forum

Date of Delivery:	30-09-2003
Title:	Evaluative review of metadata harvesting pilot
Workpackage contributing to the Document:	WP2
Total Number of Pages:	4
URL:	http://www.oaforum.org/otherfiles/tv-oai-pmh-pilot.pdf
Author:	Uwe Müller with JingYuan Wang
Contact Details:	Uwe Müller Humboldt-University Berlin, CMS (Computing Centre) Rudower Chaussee 26, 12489 Berlin, Germany

TABLE OF CONTENTS

1	Evaluative review of metadata harvesting pilot.....	2
1.1	General Approach of OAI.....	2
1.2	XML as metadata transmission structure.....	3
1.3	Dublin Core as least common denominator.....	3
1.4	OAI-PMH version 2.0 as the production protocol version	4

1 EVALUATIVE REVIEW OF METADATA HARVESTING PILOT

When the project OA-Forum officially started its work in October 2001 version 1.0 of the Open Archives Protocol for Metadata Harvesting (OAI-PMH) had already been released since more than eight months. The Computing Centre at Humboldt University had been beta tester for this first version of the OAI protocol after already the Santa Fe Convention had been implemented after its release in 1999. The change of OAI-PMH to version 1.1 in July 2001 had been of minor impact and only had to be made due to changes of the official XML specification.

Thus, the Computer Centre at Humboldt University already started with a good expertise on practical OAI experience into the project OA-Forum in October 2001.

1.1 General Approach of OAI

Generally spoken there are two principal possibilities to build up an interoperable service based on distributed metadata pools – the cross searching and the harvesting approach. While the first one rests upon a synchronous model the harvesting approach is asynchronous. When using the cross search approach for a search service the requests to the participating metadata archives are issued as an immediate consequence of a user's search enquiry. For this reason the cross search approach has the following characteristics:

- Search queries can directly be encoded (e.g. `author="Schulz"`).
- The service provider does not need a central database for the metadata.
- The service provider has to deal with duplication, ranking and merging problems synchronously.
- The service performance depends on the slowest data provider.

In contrast to the cross searching approach the harvesting approach can be described by the following properties:

- Harvesting protocols have no or only rudimentary search vocabulary.
- The service provider needs an own database to store all metadata of the participating archives. Metadata are duplicated when using the harvesting approach.
- The service provider has to regularly harvest the participating metadata archives in order to always provide the newest information. The service provider has to cope with update and deletion problems.
- Providing added value services as well as tackling duplication, ranking and merging problems is not time-critical for service providers because these processes can be conducted asynchronously.

As the name of the protocol already states the OAI-PMH pursues the harvesting approach. The crucial argument for this principal decision of the developers of the protocol which was already made during the Santa Fe meeting in 1999 was that the cross search approach does not scale for larger numbers of participating data archives. This is partially due to the fact that a service which is based on this approach always depends on the slowest participant. A strong hint for the correctness of this statement is that when using large cross search based services users first have to select a number of archives to be searched by the given enquiry. A parallel search in all archives is often impossible and at least not recommended by the service provider.

The main objective for the development of the OAI protocol was its simplicity. It should be easy to be implemented by both, the metadata archives and the harvesters who want to build up a service based on the protocol. Moreover, the protocol was designed to be based on widely accepted standards. The function of the OAI-PMH is to provide a framework for the transmission of metadata records. Delivery of complete documents is beyond the scope of OAI.

The OAI-PMH defines two groups of participants – data providers who are able to correctly reply to OAI requests and thereby deliver metadata from their local databases and service providers who makes available a service using the OAI-PMH. The protocol specification describes a set of six different request types – each with a defined set of mandatory and optional parameters. The protocol is based on HTTP, and the parameters including the request type itself can either be submitted via POST or GET parameters of an HTTP request. OAI responses always have to be encoded with XML.

The specification of the OAI-PMH comprises also defined error and exception conditions and accordant messages, a flow control mechanism to allow large archives to transmit their metadata information as a sequence of incomplete lists, and a possibility to submit general archive information such as legal or provenance information.

1.2 XML as metadata transmission structure

While the protocol altogether is based on the internet standard HTTP the OAI responses have to be encoded in XML. Therewith a further standard has been integrated into the OAI-PMH. According to general evaluations this provides a robust model for a wide network of interoperating services. Both, the request responses as whole entities and the single metadata records have to comply with well defined XML Schemas.

Alongside being a widely spread and well accepted standard which is the main reason for the existence of a growing number of tools and implementations the main advantage of XML is its extensibility. Among other things this means that the OAI protocol is suitable to transport arbitrary structured content.

For data providers who are aimed to have to overcome an even lower barrier while implementing the OAI protocol than service providers have the only necessary capability in conjunction with XML is to be able to generate well formed and – as to the given Schemas – valid XML code. Thus, data providers do not have to be able to parse and analyse XML structures. It is generally known within the OAI community that for this and other reasons a database based metadata archive can be extended by a robust and tested OAI compliant interface within several days by only one developer. A growing number of available tools diminishes even this minimal effort.

Service providers who on the other hand have to be able to parse XML code can hark back to a variety of XML libraries for almost every current programming language. Thus, XML parsers do not have to be reinvented by service provider developers. This again shows one of the main advantages of using XML as a widely accepted standard within the OAI-PMH.

1.3 Dublin Core as least common denominator

The OAI-PMH specification mandates unqualified Dublin Core as the obligatory metadata format to be delivered by every OAI compliant archive. This apparently has been constituted for reasons of interoperability. It belongs to the main characteristics of the OAI-PMH that it can serve as a metadata exchange protocol only with a defined metadata format. Thus, to allow all participants of the OAI to sensibly communicate with each other from the completion of their data provider and service provider interfaces, DC was designated to be mandatory for OAI compliant archives. Dublin Core itself is a widely accepted standard which has been established many years ago now and seemed to be adequate to fulfil this purpose.

This simple fact resulted in the most perseverative discussion within the different communities debating the manifestness of the OAI-PMH in general – including many discussions occurred within the OA-Forum project. The main objection which almost completely can be ascribed to a wrong understanding of the DC mandate within the OAI protocol was and surely still is that the usage of unqualified Dublin Core was not adequate and sufficient for many purposes. This point of criticism misconceives the fact that arbitrary metadata formats can be transmitted using the OAI-PMH. The only condition to this is the existence of an XML Schema for the respective metadata format.

Admittedly, it has to be stated that the vast majority of OAI compliant data and service providers has implemented only unqualified Dublin Core as their metadata transmission format. On the one hand this fact strongly supports the need of an obligatory standard within the OAI-PMH and retrospectively concedes the mandate of DC to the OAI developers. The assertion that the OAI-PMH owes much of its rapid distribution to the fact that Dublin Core has been made part of it can be regarded as proved. On the other hand it has to be realised that up to now the strong possibilities provided by the OAI-PMH have not been exploited to a full extend.

During the last months of the project life-span an excited discussion has been animated by the OAI inventors advocating the change of the state of Dublin Core within the OAI-PMH from mandatory to recommended. The reasons for this approach are differential. One of the arguments refers to the latent misunderstanding of the mandated Dublin Core within the OAI-PMH. Another one states that there is a significant number of data

providers having richer metadata sets within their local databases. According to this reasoning the enforced mapping to unqualified Dublin Core demolishes the incentive for data providers to offer their metadata in richer metadata formats. The most reasonable argument for the addressed approach is that some kind of data cannot sensibly be described using Dublin Core. One of the mostly used examples in this context are metadata on persons.

This topic was discussed during the most recent OA-Forum workshop in September 2003 in Bath. It turned out that many involved persons are strongly against this change of the state of Dublin Core within the protocol. Whether the approach will eventually be successful cannot be predicted within the temporal scope of the OA-Forum project.

1.4 OAI-PMH version 2.0 as the production protocol version

In June 2002 version 2.0 of the OAI Protocol for Metadata Harvesting (OAI-PMH) has been officially released to the public. Previously, the Computing Centre at Humboldt University had been alpha and beta tester of this new version and in this way contributed to the final release of the protocol. The transitional period for data providers to migrate to the new protocol version ended at the end of December 2002.

Overall, the changes between version 1.1 and 2.0 of the OAI-PMH have been quite marginal and mainly refer to recommended but not compulsory extensions. In detail the following adaptations have been accomplished:

- OAI-PMH error codes have been defined and are handled now separately from HTTP status code,
- the six requests do not any longer have their unique XML schema, all responses are now validated with one single schema,
- the protocol now supports HTTP compression functionality,
- the header element used within the response of the `ListIdentifiers`, `ListRecords` and `GetRecord` requests newly has to imply information on the set membership of the according items,
- descriptions for sets can be encoded in an XML syntax and transmitted via OAI,
- the management part of the XML response has partly been changed, e.g. the request URL of the according OAI request is now enclosed within a new XML element,
- the resumption token may contain status information on the expiration date and the total length of the complete list,
- the compulsory parameter `metadataPrefix` for the `ListIdentifiers` request allows a service provider to ask only for identifiers of those items supporting a given metadata format,
- some clarifications and restrictions have been made for date and time information,
- data providers have to inform service providers about their handling of deleted items.

The mandatory new elements, such as the error encoding and the set membership information within record headers have emphasised to be quite reasonable and useful. The more complicated functionalities, such as HTTP compression and the consistent handling of deleted items have deliberately left not required. Thus, with the new protocol version the OAI-PMH has become more functional without losing its simplicity which is the main reason for its far-reaching acceptance and distribution among digital repositories.

Version 2.0 of the OAI-PMH is designated to be a stable protocol version.