

February 2007

Institutional Repositories Workshop Strand Report

Strand title: Open Archives Protocol for Metadata Harvesting

1. Executive Summary

The Open Archives Protocol for Metadata Harvesting has been used for several years as the main mechanism to ensure interoperability between institutional repositories. It has allowed to develop a number of services. The number of repositories being developed is growing and the variety of services for scholarly material is increasing.

The OAI-PMH strands aimed to analyze the current implementations of the protocol in the IRs of the Knowledge Exchange countries, to identify the major issues they encountered, finally, to consider the necessary evolutions in the deployment of the protocol that can allow to support the new requirements of scalability and services for IRs in the next few years.

2. Summary of recommendations

- One project of national OAI knowledge bases
- Communication with the DRIVER project to coordinate the creating of data provider guidelines for IRs. Creation of mechanisms, maybe through the knowledge bases, to enforce higher degrees of compliance of KE data providers to the guidelines.
 - o Contact with the DRIVER project on the establishment of data providers guidelines
 - o Potential additions or definition of objectives for compliance levels
 - o Definition of mechanisms to encourage/enforce the guidelines in the KE countries.
- One project on persistent identifiers implementation
 - o A meeting of KE stakeholders on persistent identifiers to establish a common solution and a common strategy
 - o Creation of a test bed that would involve actors in the different countries

3. Discussion (including recommendations and items of interest)

The objectives of the OAI-PMH strand was to discuss current OAI-PMH implementations in OAI-PMH repositories, the services that are built on top of those repositories and to assess the performance and limitations of the current protocol implementations. This should allow to identify issues related either to the protocol or to its implementations.

A number of topics have been identified in the briefing paper on OAI data providers interfaces, access to IR digital documents from metadata conveyed through OAI-PMH and related services. A questionnaire has been circulated. It covered quality issues, digital objects, services, tools for data and service providers, finally marketing issues. A compilation was made from six questionnaires. Those allowed to identify a number of important issues to be discussed over the meeting.

The discussion focused on analyzing the current issues with OAI implementations and anticipating its evolutions for the future. The current framework will eventually have to adapt to different usage models and to scale up with the multiplication of IRs.

3.1 A set of problems have been identified during the discussion

Stability of the protocol

The protocol was considered very performant but implementations are often not very reliable in practice. The availability of the data provider service for example is often not guaranteed. There is no good mechanism to inform service providers that the repository is temporarily out of order. Moreover, simple errors can create problems in the harvesting process. While many data providers do not keep track of who harvests them the communication of information is a little difficult.

A major difficulty is that for data providers, the protocol is “silent”, that means a data provider does not know when a harvest has failed or when there are problems with its metadata unless a service provider tells it. Service providers on the other hand do not have a standard mechanism to communicate about the information they have on specific data providers. Some types of information could however even be read by service providers automatically.

OAI sets

OAI Sets are used to arrange a repository for different aspects or to harvest big repositories in multiple chunks. Some software packages have predefined sets. But they all use different criteria and refer to different classifications if any (for eg. LC Class in the UK and DDC in Germany for topic-related classification in OAI sets). Normally the classification is not used in its total depth.

There is no common structure of repositories and set names are not always informative. When there are too many sets, this may mean that the sets are trying to solve a problem that should be solved another way. Finally, there is no mechanism for service providers to harvest records that belong to two defined sets for example.

Sets do not have an update mechanism like records, there is no mechanism to show that a set was deleted, that resources have been moved out of a set to another set. Registries could have a role in taking picture of the set organization and documenting its evolution.

Deletion strategies

Many repositories choose for the deletion strategy ‘no’, that means they do not support a mechanism (transient or persistent) to convey the information to service providers that a record has been deleted. This requires the service providers to reharvest the full repository fairly often instead of performing incremental harvests (changes to the repository since their last visit). This is a major constraint essentially for large repositories that can take several days to harvest.

In order for the harvesting mechanisms to generalize, the OAI-PMH interactions need to scale up. The adoption of a deletion strategy else than ‘no’ should be recommended.

Metadata format

OAI_DC is the only mandatory metadata format at the moment. It does not allow to express unambiguously a number of information that would be useful to develop new services in the future. As a result, many repositories (at least in Denmark, Germany and the Netherlands) also expose one or more alternative formats. But those vary in a broad way. It is difficult for a service provider to develop reprocessing streamlines for a large variety of formats. The analysis of the formats used by data providers as alternatives to Dublin Core do not show a consistent alternative to OAI_DC. One solution could be to adopt an alternative common format that would be made mandatory additionally to simple DC. Candidates could be DIDL XML containers or the ePrints application profile. Another option could be to deliver the original format, for example the format of the database behind. A useful resource is the metadata format registry at UKOLN.

Datestamp granularity

This topic was not identified as major problem. The datestamp of metadata records is however not modified consistently by data providers when there is a real update of the content of the metadata record.

Metadata XML parser errors

Invalid xml files are encountered in about 10% of the repositories, particularly in DSpace implementations. Further analysis should be performed on the technical reasons for this (eg. could be because rely on earlier Java versions). Usually email communication is sufficient to repair the problem. But the distribution of responsibilities and competences is often uncertain, the person in charge of the repository does not always have directly the possibility to make the appropriate modifications.

Documentation of data providers implementations

Data providers have the opportunity of documenting their repository using the OAI Identify and ListSets responses. Default containers have been implemented in different software packages (ePrints description schema). But many times, the data providers do not add to them or even do not modify default values. When information is changing, they do not record that change (eg. administrator email contact). OpenDOAR has created a generator of policy descriptions for ePrints repositories. People need a model for this, rather than the indication of a schema or even a simple form. Software packages should also implement more consistently default containers for encouraging data providers to deliver better descriptions of their material and implementation. The major difficulty is that structure of documentation is modular (one schema for labels, another for ePrints policy description, others for collection description, another for IPR specifications). A standard set of information should be easier to use for data providers. Friends networks could increase with Web 2.0 as a discovery tool for repositories. Registries should also use those documentation information (harvest them) and make them available as a comprehensive documentation of the data providers.

Software packages

A number of implementation issues are related to default configurations and capabilities of software packages. It is therefore very important to establish contacts with the major software package developers. User communities are very active, particularly Opus and ePrints are part of the KE countries. It is less certain however that developers of DSpace and ePrints for example communicate. Fedora is trying to be a model JSR implementation.

Any type of guidelines have to be coordinated with default implementations of major software packages, that is Dspace, ePrints, Opus, Arno and Fedora.

Resource harvesting, Pointers and Complex Objects representations

There is no mechanism in the OAI-PMH protocol to harvest the actual resource (digital object). However, in order to build several services, it is possible to access the digital object (eg. Full text indexing of articles, overlay journals evoked in Germany).

The DC records do not allow to consistently represent complex objects (typically objects that cannot be addressed through a single pointer).

The DARE network has developed guidelines for the usage of an XML container (DIDL) that embeds both metadata records and pointers to multiple representations and/or parts of digital objects. However they are not metadata formats but containers that need application profiles to allow interoperability.

But there are differences in access policy in different institutions. Several institutions may not consider favourably the possibility for an external service to harvest digital objects. There should be a mechanism to specify a remote copy policy. Some objects also have different copyrights applying to different parts of the object (eg. A thesis). The development of a mechanism that would allow services to reliably use pointers to digital objects or parts of digital objects should include ways of specifying access policy considerations.

The OAI-ORE initiative is currently dealing with object access and harvesting. It seems to be discussed mainly in the Netherlands and less in Denmark, Germany and the U.K.

Open Access / Licensed material

IR do not only contain open access digital objects. Some objects can even be under embargo. There is no way for service provider to know a priori that a digital object is available in open access.

There should be a mechanism to distinguish between metadata describing open access material, metadata describing analog only resources and, metadata describing non open access material.

OAI sets can allow data providers to specify that distinction in their collections. Context Objects for the resources can then make the access distinction resolvable for service providers to indicate that information to their users.

Central platforms

When there is no possibility or interest for an institution to implement its own IR, it can or should use a federation of repositories or “central platform”. They are usually more stable but they complicate the contact with individual providers themselves. It is a transitional solution that may have to be developed while the IR phenomenon scales up.

Aggregators

Aggregators harvest from multiple IRs and re-expose the metadata, possibly after reprocessing them. A service can then harvest from the aggregator instead of from individual IRs. This is more reliable for service providers. But the data is less fresh.

Intute for example came out of the ePrints project. The idea was to reprocess then sent back the data to the source IR.

Aggregators can also be a solution to the preservation of information in case of failure or instability of an IR. It could make sense for each country to have a national aggregator.

Registries

Existing registries are very useful tools. They would deserve to be developed with more information about data providers. They should take advantage of different information sources (manually recorded by data providers, collected by the registry manager, automatically harvested from the repositories and recorded by service providers).

Registries can be used as way of improving the process of harvesting for example but they will not represent the only way to cover information about a repository. Since there are different players which are interested in such information it makes no sense to address a specific proprietary registry interface but to put the focus on a transparent communication protocol as OAI PMH.

While the OAI-based infrastructure expands, registries can have a prominent role, for resources discovery as well as for facilitating the communication between data and service providers.

Again a national structure for gathering information from different sources which redistributes this information as a knowledge node would make sense.

3.2 Two major issues to consider

The discussion has allowed to identify the following issues as major challenges for OAI implementations:

1- **Communication between data and service providers on technical implementations and content related issues.**

This includes the role of registries for data providers discovery, the way in which data providers provide information on their collections, the place of registries in the harvesting process. The communication between data providers and service providers as well as between service providers themselves is seen as an important factor to improve the stability of OAI-PMH implementations.

2- **Access to digital objects contained in IRs**

This includes two major issues:

- the identification by service providers of metadata records relating to digital objects freely accessible, those relating to digital objects with access restrictions and those relating to analog only objects
- access to the digital objects from the metadata record. The OAI-PMH protocol is essentially used to convey metadata rather than content (DIDL containers only contain pointers in the DARE implementations). The link to the digital object from the metadata records is done through one or more pointers. The pointer eventually uses a unique and persistent identifier. The way in which pointers are recorded should allow service providers to know for sure how to get to the digital object or to a jump-off page etc. It should also allow to point to complex objects.

4. Outcomes

The reason why the protocol works successfully is because it is simple and does not try to solve all the problems. It is a standard that allows to convey structured information. An extension of the protocol would necessarily be an additional layer of complication. Major evolutions should therefore reside on the one part in reinforcing the infrastructure that support the OAI framework of communication between data and service providers and the digital objects repositories, on the hand to establish common guidelines in the Knowledge Exchange communities.

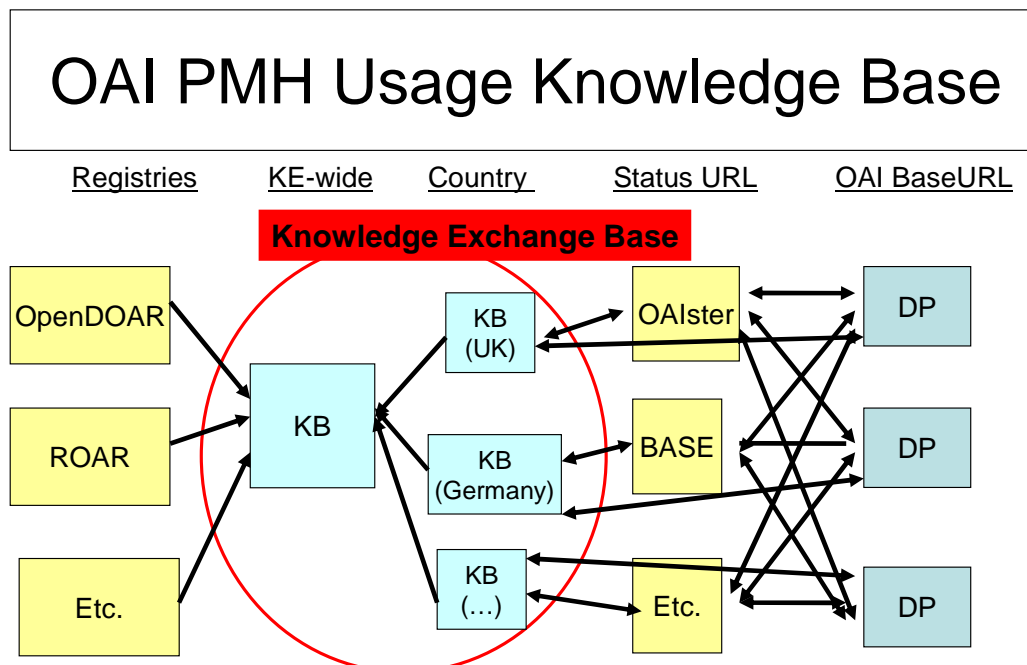
4.1 Communication between data and service providers on technical implementations and content related issues.

To improve the current situation of OAI PMH usage among institutional repositories a national body in each KE country could be established which could be in charge for

- discovering OAI providers in their country, (and neighbourhood countries)
- acting as a connector between data providers and service providers
- collecting detailed information on repositories and their characteristics (manual and automatic checks)
- informing data providers about recommendations and problems with their service
- distributing the collected information (user web interface, technical interface)

The architecture should serve as an information exchange platform and thereby improve quality and visibility of the addressed repositories. In order to build an open interface this strategy intends on the definition of a communication protocol between the players (Service description). Extra xml schema(s) (as extension to OAI) covering metadata about the protocol usage could be a possible solution. In order to accomplish the national information a KE-wide consolidation of these efforts should take place.

The following figure shows the designed infrastructure and its information flow:



The concept includes

- communication with a KE group for protocol and schema recommendations
- Establishing national Knowledge base infrastructure (using existing resources)

- Establishing an KE-wide aggregation of the national knowledge bases

These activities should be coordinated with the DRIVER project and can be achieved in the next 12 – 24 months.

The collected information for the knowledge base includes in detail

- Extended description of the IRs
- Entry page
- Official names
- Official icon
- System platform (including version)
- Reliability incl. current status
- Harvest speed
- Invalid XML format
- Format problems (Non-utf8, no urls, (neither creator nor title)
- Format specifics
- Size including a logging of the harvest processes
- Resumption token usage
- Harvest problem (expiration date)
- Incremental Harvesting. support
- Granularity usage
- Set usage
- Metadata Format specifics
- Fulltext marking
- Fulltext percentage
- Classification/Subject usage

4.2 A set of guidelines for data providers

A number of difficulties related to access to the digital objects can be overcome by using implementation guidelines.

1- On repositories organization

Sets can be created for analog only resources, digital objects in open access and digital object in restricted access when an IR does not only contain open access digital resources. A service provider could therefore harvest the sets it can handle. If a service provider is interested in digital objects, some of those objects may or may not have access restrictions for different users of the service.

A naming convention could be adopted in order to identify the different criteria adopted for classifying sets, for example, the usage of a hierarchical structure such as `Access:openaccess`, `access:analog` and `access:restrictedaccess`, `class:lcc:xxx`, `class:ddc:yyy`. This would allow a better reusability of sets by service providers, not only for selective harvests but also for the usage of the subsets of information as organized by the data providers.

Finally, a mechanism to request a combination of multiple sets (eg. Items that belong to `Access:openaccess` and to `class:lcc:xxx`) could be investigated.

2- On links to resources

No usage scenario was found convincing to extend the protocol to allow resource harvesting at this point.

A solution to the current usage scenario was adopted in the scope of the DARE project: XML containers (DIDL AP) to convey complex objects and links between metadata and object. It allows notably to access different parts of complex objects and to specify what links contained in the record actually point to (eg. a jump off page vs a digital object in PDF format etc). A mechanism that links different object parts could be investigated, as well as a mechanism that would allow to link to external objects.

Besides that a mechanism should be investigated to allow to rebuild a ContextObject (OpenURL) to the resource. A service provider could therefore adapt the display of resources for example to the individual cases of users.

In order to make sure that the guidelines can be implemented, it is necessary to communicate with major software developers, either through user communities or by direct contact with the developers' organizations. Software packages can include standard documentation mechanisms, default OAI sets etc. Need to work with other entities that create guidelines for OAI data providers for IRs, notably the European project DRIVER. There is an overlap in the people involved in the DRIVER project and in the Knowledge Exchange community. Contact should be made over the next 6 months to make sure that those considerations are taken into account in the DRIVER implementation guidelines. If some of those remain out of the scope of the DRIVER concern, they could be thought as a possible strategic orientation for IRs in the Knowledge Exchange countries, notably through the creation of dedicated services that would make use of each of those features.

4.3 Adopting a common approach to persistent identifiers

A limitation of the current linking mechanisms is the absence of universal and persistent identification model. The Knowledge Exchange Forum could be a good area to implement a common solution that would take into account the experiences of different partner countries.

The current major schemes are notably URNs, DOIs and ARKs. They have been considered in the scope of a SURF sponsored study. It would be possible to organize a meeting of the people involved in the major initiatives launched in the different countries within the next 6 months. A common solution could be adopted by the different partners. If appropriate, national resolver implementations could be decided. An agreement could be found for replication and redirection of pointers for different identifiers between the Knowledge Exchange partners.

5. Annexes (including supporting papers, referenced materials, list of participants)

Muriel Foulonneau muriel.foulonneau@ccsd.cnrs.fr
Friedrich Summann friedrich.summann@uni-bielefeld.de.
Jochen Schirrwagen schirrwagen@hbz-nrw.de
Paul Walk p.walk@ukoln.ac.uk
Dr. Peter Millington Peter.Millington@nottingham.ac.uk
Drs. Laurents Sesink laurents.sesink@dans.knaw.nl
Maarten Steenhuis m.m.a.j.steenhuis@library.leidenuniv.nl
Kasper Løvschall kl@aub.aau.dk
Franck Falcoz franck@cvt.dk

DFG

JISC

DEff

SURF