ORI-OAI: Referencing and Indexing Tool for a Network of OAI Portals

Raymond Bourges[1], Yohan Colmant[2], François Jannin[3]

[1] Université de Rennes 1, France - [2] Université de Valenciennes, France - [3] INP ENSEEIHT Toulouse, France.

raymond.bourges@univ-rennes1.fr francois.jannin@enseeiht.fr
yohan.colmant@univ-valenciennes.fr

1 Introduction

The ORI-OAI project [1] seeks to create an open system, one that is both open source and royalty free. This system allows users to:

- manage all the digital resources produced by universities,
- share these resources with other institutions of higher education,
- valorize these resources with high-quality indexing, and
- make these resources e-accessible via ergonomic interfaces, according to well-defined access rules.

Clearly, the exploding production of digital institutional documents by/for universities has led institutional stakeholders to consider the need for a comprehensive, coordinated management schema for this production in order to insure access over time. Many of these stakeholders (e.g., teachers, researchers, students, centralized documentation services, TICE[1] services) produce digital documents, while others (e.g., centralized documentation services, IT systems management) work to valorize and manage their contents over time. The development of Thematic Digital Universities (TDU) and Regional Digital Universities (RDU) has led to questions about how these resources can be shared in interoperable systems accessible both through the institutions' digital workspace[2] and through the TDU and RDU.

1.1 A brief history

The ORI-OAI project is publicly funded, primarily by the Libraries sub-division of the Higher Education Administration and the TICE sub-division of the French Ministry of Education's Technology Administration. The ORI-OAI system is the fruit of our deliberations concerning several digital resource management initiatives:

- SYNAPSE (INSA; Lyon, France)
- inJAC (ESUP-Portail consortium; France), and
- the pedagogical resources portal of the TDU UNIT consortium (France).

2 Functional Outline

2.1 The stakeholders

The ORI-OAI project is intended to allow users to search for digital resources from an institution or a group of institutions, and then give those users restricted or unrestricted access to those resources. In addition to facilitating user access to these resources, the system is designed to allow resources to be posted on the system by as many people as possible, including teachers, researchers, administrative staff, and even students. The more complex indexation operations are left to specialized services, such as the centralized documentation service. In general, the system provides a high level of configuration flexibility so that

[1] TICE (*Technologies de l'Information et de la Communication pour l'Education*) Information and Communication Technologies for Education

[2] ENT (*Espace Numérique de Travail*), the web portal providing access to university services in conjunction with the institution's IT service.

institutions can choose their own organization for resource posting, referencing and indexing procedures.

2.2 The digital resources

The ORI-OAI system was, from its inception, designed to work on different types of digital resources. All resource types can be referenced by using an appropriate metadata format. The referenced resources can be pedagogical resources (e.g., course photocopies, problem statements, corrected exercises), student work (e.g., internship reports, projects, bibliographic summaries), research (e.g., publications, technical reports, Master's theses and doctoral dissertations, and/or published technical documentation acquired by the institution (e.g., e-periodicals, e-books), for example.

2.3 The functionalities

The ORI-OAI system was developed in order to respond to several objectives. Some of the key elements are listed below:

A unique reference framework for the institution's digital resources: This reference framework does not replace the diverse platforms that can use and publish these resources (e.g., pedagogical platforms, laboratory websites), but it does manage the standard form of the resource and can allow access controls within an identity federation.

An advanced multi-criteria search system that allows different fields of metadata to be searched.

Thematic access to resources, according to simplified classifications based on the Dewey decimal system.

A digital resource management and publication system via web publication with access rights supervision. Resource descriptions based on such norms as LOM 5, TEF 5, and Dublin Core 5, for example, in conjunction with other documentation systems for the sharing of authority tables (e.g., SUDOC 5, STAR 5); Indexation according to standard university library classification systems (e.g., the Dewey decimal system), and exploited via specific TDU classifications; Digital resource archiving.

A production system implicating the involved stakeholders in the procedures developed and materialized by workflows; Management of document versions and access to the native versions of these documents by those that created them.

A document sharing system based on exchanging metadata according to the OAI-PMH 5 protocol, which allows the system to function as part of a community (e.g., TDU) using a portal network.

An open-source system that is royalty-free, documented and easy to install.

3 Concepts

3.1 The metadata concept

Metadata represent the set of external information associated with a resource. This information set allows a use context to be defined for each resource, as well as a set of inter-resource relationships. For example, an author name can be used to find all the resources produced by that person, or the name of a taxonomic code can be used to obtain a semantically-structured vision of a group of resources.

Through concrete formats based on real-life practice (e.g., Dublin Core, LOM, CDM 5, TEF, MARC 5, ETDMS 5), the concept of metadata defines a fixed group of data that can be semi-structured using the XML markup language. This group aggregates and anticipates the needs of the community around the sum of its activities. In addition, it provides a strong semantic component due to the use of standardized closed vocabularies, which, thanks to the quality of the information thus produced, make the system highly efficient. These closed vocabulary sets are defined inside XML schemas which validate the metadata for a given format.

A data repository is a system that is able to associate one or several of the documents in a metadata set in order to constitute a reference framework for an open set of utilizations, within the limits defined by the format used. Thus, the non-qualified Dublin Core format defines a set of 15 very general, optional and repeatable metadata elements, which may be enough to represent simple non-ambiguous relationships in a resource set: a Master's thesis can usually be described by its author's name, the date of creation and publication, and a code situating the resource in the domains set of the school where it was produced, as well as a few other metadata elements.

More complex, the LOM format conceptualizes all the pertinent information describing a pedagogical object, its rich structure providing details about each of the aspects that allow the object to be apprehended within its activity set: for example, the pedagogical aspect for school-based management and the pedagogical conditions that determine the audience (e.g., level, type of end user, duration), the legal aspect for the legal conditions, and the lifecycle aspect for the object's evolution and versions. This format can even define a meta-metadata aspect for describing information about the metadata themselves.

In summary, by describing each resource appropriately, the metadata serve as the foundation of a quality referencing system that gives each resource its place in the network of relationships that compose an activity, which allows those involved in this activity, both as producers and consumers, to locate and manipulate each resource as easily as possible.

3.2 The OAI-PMH protocol

The OAI-PMH 5 protocol was created by the organization, Open Archive Initiative 5, which, as its name implies, is dedicated to open archives and resource sharing. Each resource is, thus, represented by an OAI **identifier** composed of an OAI header and a metadata set (cf section 3.1). The heading contains the information used by the OAI protocol for the selection criteria, such as the identifier modification date, the sets to which the identifier belongs, or even the identifier's cancellation status. Each metadata format must be described by an XSD (i.e., an XML Schema Definition), whose name and internet file address are defined in the OAI responses, which makes any format described by an XSD as eligible as any OAI format.

Using a simple but effective syntax composed of 6 verbs, OAI-PMH allows these identifiers to be distributed by a mechanism that calls upon the **harvester**, which is a service provider, and one or several **repositories**, each one associated to a data provider.

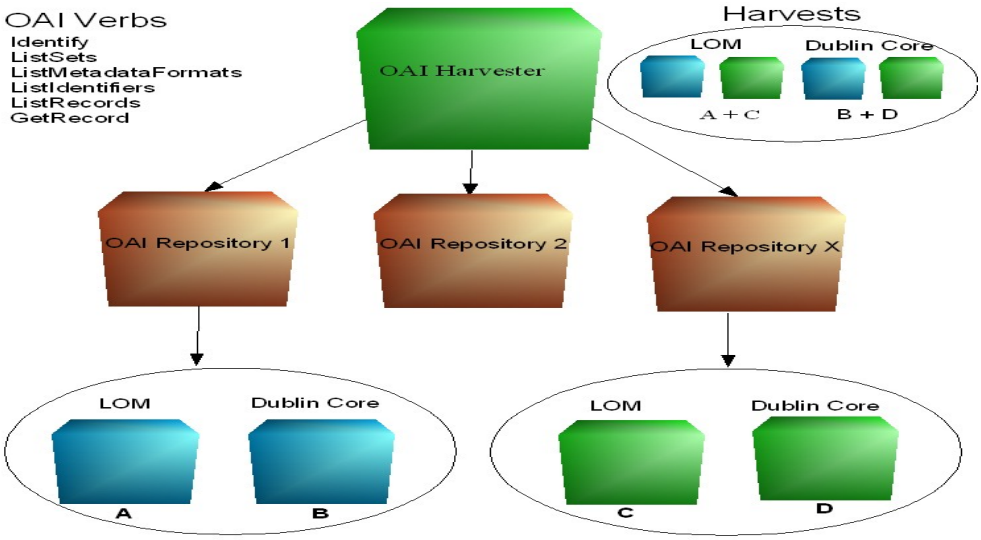

Figure 1 – *OAI schema*

As shown in Figure 1, the general mechanism calls upon an OAI harvester, which selectively harvests the records stored in several OAI repositories, according to different **selection criteria**. A harvester's selection criteria must define at least the desired metadata **format type** (e.g., LOM or Dublin Code). Optionally, the criteria can define the starting and ending **dates** and a **set** targeting a domain, or any form of appropriate group for a specific use, more precisely. The notion of set is an internal function of the repository; this function groups together all the identifiers whose common metadata formats have elements that can be classified. The date criteria make it possible to perform incremental harvesting (i.e., harvesting only the identifiers that have been added, modified or cancelled since the last harvest), which makes it easier for the harvester to correctly process the lifecycle of the harvested identifiers.

The first verb is ***Identify***, which provides a repository identity record that may contain a description of the available content or the links to other "buddy" repositories. The verbs, ***ListSets*** and ***ListMetadataFormats***, respectively provide a list of the available sets and a list of the formats supported. Access to the records is possible using the verbs, ***ListRecords*** and ***ListIdentifers***, the second sending only the heading and the appropriate identifier. The identifier makes it possible to access a single record individually using the sixth verb, ***GetRecord***.

The OAI verbs that return lists use a resumption token to cut the lists that are too long into smaller segments, which can be more easily manipulated in the network flows.

3.3 A Shibboleth identity federation

Identity federations seek to facilitate on-line digital resource sharing between institutions by interconnecting the institutions' authentication services. This interconnection makes it possible to open access to a resource (e.g., pedagogical, scientific, editorial, business application) to an identified population, without needing to manage user registration locally.

The identity federation concretizes, for a group of institutions, the interconnection of the authentication services and the use of a common set of user attributes. An institution that manages a user set is called an identity provider. A service provider is an entity—institution, public administration, private company—that proposes a digital resource on-line as part of the federation. A single institution can participate in several federations and manage partnerships bilaterally. This institution can also play both the role of identity provider and service provider. Technically, the trust relationships between members of a federation are based on

electronic certificates and the shared definition of the different identity and service providers. In addition, trust can be established between federation participants through a formal convention.

In French institutions of higher education, the identity federations use rules defined by the CRU[3] 5, which also provides technical assistance. The implementation technique chosen by both the CRU and the ORI-OAI project is the one developed for the university Internet2 context: Shibboleth 5.

The following schema shows a simplified version of the mechanism used by the identity federation:

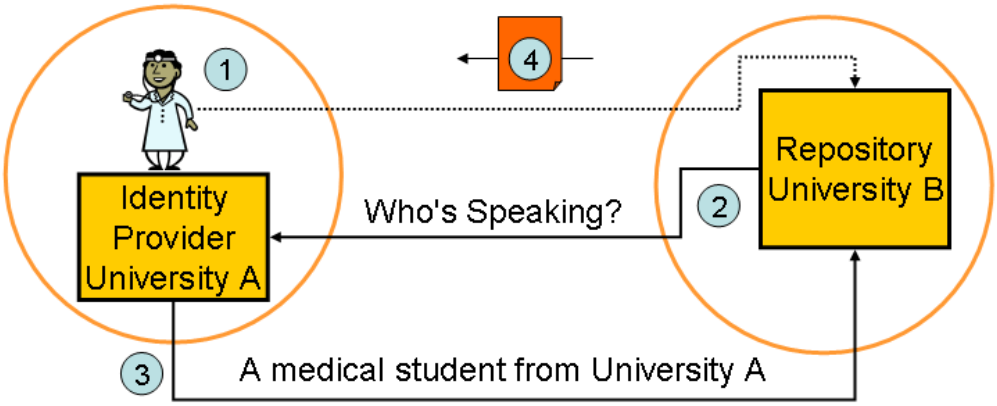

Figure 2 – *Shibboleth mechanism*

1. The user attempts to access a document in another institution.
2. Institution B requests information about the user who is transmitting the document request from the identity provider of institution A, to which to the user belongs.
3. Institution A provides the user's attributes to the Institution B service provider.
4. The Institution B service provider verifies these attributes in order to decide whether or not to send the document.

3.4 Examples of application contexts

A harvester assembles, in a single location, a set of metadata from different repositories. It thus becomes possible to index and propose a search interface for these metadata in order to locate and provide access to the associated resource.

Numerous applications can be envisioned, two of which are described briefly below:

- A person surfing the internet consults a consortium website (e.g., a TDU or a RDU) to look for resources that are *a priori* royalty free.
- An institutional user, within the framework of the services proposed by the institution's digital workspace, searches for resources in the his/her own institution and/or partner institutions. In this case, the resources could be subject to access control.

[3] CRU (Comité Réseau des Universités) : Technical support organization for French Universties

3.4.1 Access through a consortium website

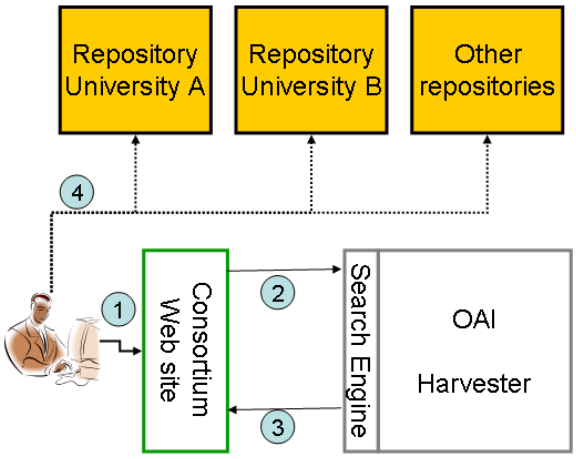

Figure 3 – *website context*

1. The user accesses the consortium's website.
2. He/she activates the consortium's search engine.
3. He/she obtains references for certain resources.
4. He/she accesses these resources in the different repositories of the consortium's member institutions.

3.4.2 Access through an institution's digital workspace

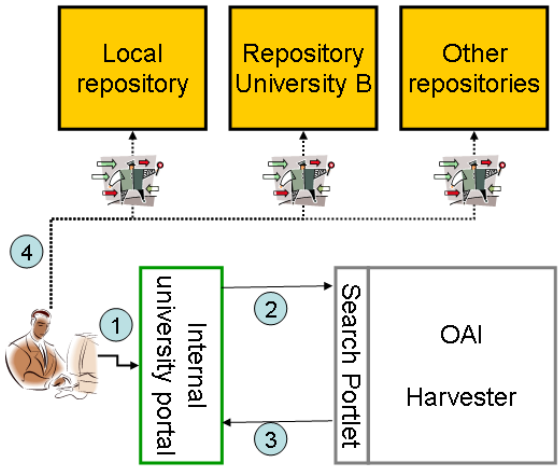

Figure 4 – *digital workspace context*

1. The user identifies him/herself in order to access the institution's service portal.
2. He/she formulates a search request on the ORI-OAI search portlet *(JSR 168)* 5, installed in the institution portal.
3. He/she obtains references for certain resources.
4. He/she accesses these resources in the repositories of his/her own institution and/or the partner institutions. If the resources accessed require an access control, the identity federation's mechanisms are used; a transparent single sign on (SSO) means the user doesn't need to re-identify him/herself.

4 Implementation

4.1 The general architecture of the ORI-OAI project

The architecture of the ORI-OAI project is composed of 7 independently developed modules. Each of these modules plays a well-defined role in the system and communicates with the others through services revealed on the front-end processor of each module. The technology chosen to make dialogue between the modules possible is the Web Services technology 5. This choice provides a great deal of flexibility in the architectural variations and the programming languages that can be used. This architectural technology was designed to permit external softwares to dialogue with the various system elements. For example, this technology makes it possible for a library to search for documents using its documentation portal without passing through the search interface, ORI-OAI-search, by establishing a direct connection between the library's software and the search index, ORI-OAI-indexing.

Please note that ORI-OAI was developed as Open Source under a General Public License 5.

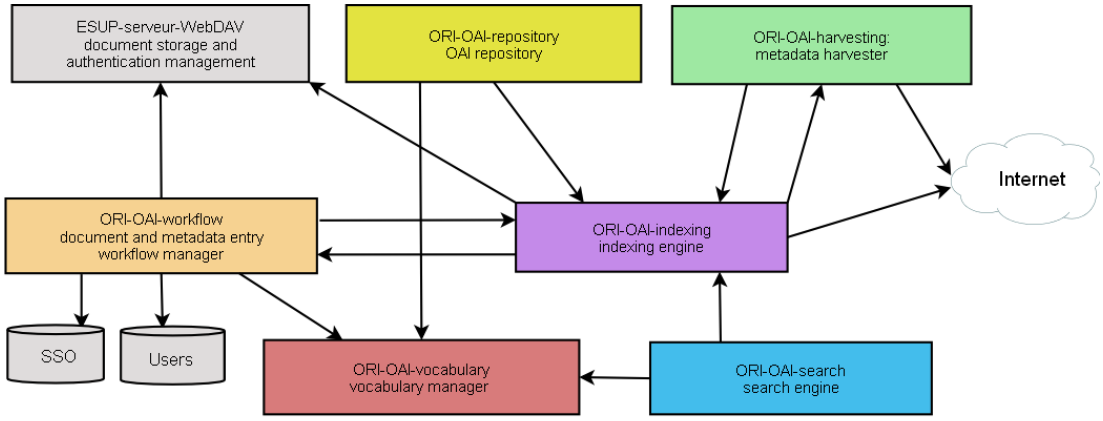

Figure 5 – *The general architecture*

4.2 The different modules

As mentioned above, all the modules were developed independently. However, they all are based on certain common technologies: for example, the development language Java 5 and the development framework Spring 5 are used in each module and all Web Services communication is implemented through XFire 5.

The following paragraphs present the functionalities of each module as well as the technologies that are employed.

ORI-OAI-workflow

The ORI-OAI-workflow module is used to place documents in the system and to assign resource access rights, as well as to enter metadata. This placement engine uses a variety of technologies that permit complete, highly precise and powerful parametering appropriate to the needs of the institution or the consortium that sets up the system.

The workflow engine, OsWorkflow 5, allows the detailed configuration of all the actions carried out during each of the placement steps. Such configurations make it possible to respond both to the needs of institutions requiring the person placing a document in the system to complete all the metadata fields, and the needs of more structured institutions for whom the metadata is entered in several steps by different people. Each institution can this configure the module to best respond to its individual needs.

The metadata entry forms can also be configured to respond to institutional needs. The XForms 5 technology associated with the transformation engine Orbeon 5 provides dynamic data entry forms based on configuration files written in XML. The support for the different metadata formats is based on the possibility of enriching the ORI-OAI-workflow module using new Xforms configurations.

The other technologies used in this module are Acegi Security 5, which manages all the application's access controls; eXist 5, which provides an Open Source XML database for storing the metadata records; Hibernate 5, which provides transparent access to the relational database; and JSF 5, which manages the graphic interfaces.

ORI-OAI-indexing

Once the resources have been placed in the system and the entered metadata has been validated, the resources are then indexed by the ORI-OAI-indexing module, whose job it is to index the metadata record as well as the associated documents. To do so, the module uses the

Lucene indexing engine 5. Capable of indexing different sources, Lucene provides a fast, powerful search engine based on diverse analyzers. For example, the French language analyzer is able to handle conjugated verbs and plurals, as well as accents and special characters. A weighting system can be used to give one metadata field more importance than another, for example, putting the documents whose title contains the search element higher in the results list than those in which the search element is found in the description.

Lius 5 is an indexing framework based on the Jakarta Lucene project. It permits the indexing of different file formats, including XML, PDF, OpenOffice, ZIP files and MP3. It is used in our project to allow the advanced configuration of the index fields in the different formats of the XML metadata records and the indexing of the associated text documents. In addition to indexing, ORI-OAI-indexing uses the syntax of Lucene requests in a document search service via Web Services. This search service is used by the different system modules.

ORI-OAI-repository

The OAI-PMH protocol (§ section 3.2) is used by the ORI-OAI-repository module to expose the metadata records entered in the ORI-OAI-workflow module. ORI-OAI-repository uses the software OAICat 5 to expose the records so that they can be harvested by any OAI harvester.

This module also manages the OAI protocol concept of "OAI-PMH sets", which allows the metadata records to be exposed as distinct sets, often connected to a specific theme. Using the OAI-PMH sets concept makes it possible, for example, to identify the set of all the pedagogical documents in the LOM format that deal with mathematical notions. To identify the documents corresponding to the different sets, the ORI-OAI-repository module builds requests based on the criteria associated with the different sets and sends them to the ORI-OAI-indexing module.

ORI-OAI-harvesting

This system module corresponds to the OAI-PMH harvester. Using the software, OAIHarvester2 5, the ORI-OAI-harvesting module allows the metadata records to be harvested in all OAI repositories. Via a user-friendly graphic interface, the system administrator can program the different harvests that will be launched using the task manager, Quartz 5. The records harvested are then stored locally in the XML database, eXist. Like the workflow engine, this module sends all the metadata records harvested to the indexing engine to be indexed. The Lucene search index thus contains local documents, but also the metadata records harvested by ORI-OAI-harvesting.

ORI-OAI-search

The ORI-OAI-search module provides a graphic interface for in-system document searches. By dialoging with the ORI-OAI-indexing module, this search module generates requests in Lucene syntax and lists the documents found. This module allows total configuration freedom both in terms of the document formats sought and the types of searches proposed to the user.

There are three possible search types:

- By date: the user may search for documents by date of creation or modification, for example, which is useful for seeking out new additions to the database.
- By criteria (simple or advanced): the user may configure different search forms, including a range of criteria from simple to advanced. A simple search form might include only one field, thus allowing a search for a complete document and its associated metadata; a more advanced search would involve a search form that allows an independent search of each metadata element in a given format.

- By subject: in this type of search, the user is not asked to enter any information, but rather to choose from the proposed elements. This type of search allows users to search by document classification, author, or a keyword, for example. It may be requested by users who would like to discover the documents referenced in the system without using specific search criteria. This type of search can also guide users in their search by proposing only the values really indexed, such as in the case of the keyword search.

Please note that this module exists in two versions: *servlet* for standard installation on a web server, and *portlet* for integration into a digital workspace (ENT).

ORI-OAI-vocabulary

The ORI-OAI-vocabulary module is the one that manages the vocabularies used by the different modules. "Vocabulary" is used here to refer to the closed set of values available for a given criteria. These vocabularies can be static and configured with XML files: for example, document classifications or strict values of the LOM metadata fields. They can also be dynamic, as in the case of the values available in a specific metadata index. For example, a list of the key words or authors already entered in the pedagogical document database can be reconstituted dynamically by querying the ORI-OAI-indexing module.

These vocabularies are used by:

- The ORI-OAI-workflow module to propose lists of values during the metadata entry procedure.
- The search engine, ORI-OAI-search, to search by subject or by available values for certain fields in an advanced search.
- The repository ORI-OAI-repository to dynamically generate OAI sets, for example, for a given subject.

ESUP-server-WebDAV

Initially developed by the ESUP-Portail consortium, this WebDAV server can store documents on-line using the WebDAV protocol 5. This module is based on the Jakarta/Slide server 5. Using the different methods in the WebDAV protocol, it is possible to manage distant documents (e.g., placement, downloading, cancellation). In addition, ACP (Access Control Protocol) allows user document access rights (e.g., read only, read/write) to be managed dynamically.

Using this server as a foundation, the consortium has developed several elements to complete it:

- Authentication support using SSO (Single Sign-On) CAS 5
- Identity federation management using Shibboleth
- Outsourcing of user group management
- Quota support

5 Conclusion

Currently, version 1 of the ORI-OAI project uses the Dublin Core metadata set and the LOM standard to provide a referencing tool for pedagogical resources as well as resources connected to scientific research (e.g., Master's theses, scientific publications). With the full implementation of the OAI-PMH protocol and the use of Web Services, this tool lays the foundation for widespread interoperability with other systems. Its workflow management system permits a flexible configuration of the roles of the various system stakeholders during

production and validation of metadata, whose quality is reinforced by the use of standardized vocabularies shared by the different system modules.

The modular composition of ORI-OAI thus forms a functional core that was intentionally designed to be quite open to numerous alternative uses, making interoperability possible with national systems (e.g., HAL 5 and STAR), as well as other applications that are widely used in the French higher education community (e.g., the pedagogical platform, Moodle5).

The hard work of the French Education and Research Ministry and the various TDU and RDU involved in the project is now paying off and the project is beginning to proliferate. The goal is the widespread adoption of this tool by the majority of higher education institutions in France, and perhaps beyond via the international partnerships that unite universities in countries around the world.

Bibliography

[1] http://www.ori-oai.org/

[2] http://www.esup-portail.org

[3] http://www.oclc.org/dewey

[4] http://ltsc.ieee.org/xsd/lomv1.0/lom.xsd

[5] http://www.abes.fr/abes/documents/tef/index.html

[6] http://dublincore.org

[7] http://www.abes.fr/abes/page,352,le-reseau-sudoc.html

[8] http://www.abes.fr/abes/page,428,star.html

[9] http://www.openarchives.org/OAI/openarchivesprotocol.html

[10] http://cdm.utdanning.no/cdm

[11] http://www.loc.gov/marc

[12] http://www.ndltd.org/standards/metadata/etd-ms-v1.01.html

[13] http://www.openarchives.org

[14] http://federation.cru.fr

[15] http://shibboleth.internet2.edu/

[16] http://jcp.org/en/jsr/detail?id=168

[17] http://www.w3.org/2002/ws/

[18] http://www.gnu.org/licenses/gpl.html

[19] http://www.java.com/

[20] http://www.springframework.org/

[21] http://xfire.codehaus.org/

[22] http://www.opensymphony.com/osworkflow

[23] http://www.w3.org/MarkUp/Forms/

[24] http://www.orbeon.com/

[25] http://www.acegisecurity.org/

[26] http://exist.sourceforge.net/

[27] http://www.hibernate.org/

[28] http://java.sun.com/javaee/javaserverfaces/

[29] http://lucene.apache.org/java/docs/index.html

[30] http://www.bibl.ulaval.ca/lius/

[31] http://www.oclc.org/research/software/oai/cat.htm

[32] http://www.oclc.org/research/software/oai/harvester2.htm

[33] http://www.opensymphony.com/quartz/

[34] http://www.webdav.org/

[35] http://jakarta.apache.org/slide/

[36] http://www.ja-sig.org/products/cas/

[37] http://hal.archives-ouvertes.fr

[38] http://moodle.org