ORI-OAI

Outil de Référencement et d'Indexation pour un réseau de portails OAI-PMH

Yohan Colmant
Service Informatique - Université de Valenciennes
yohan.colmant@univ-valenciennes.fr

Nolwen Huet
Doc'INSA - INSA de Lyon
nolwen.huet@insa-lyon.fr

1 Introduction

Le projet ORI-OAI [1] vise la mise en place d'un système ouvert, en open source, libre, permettant :

- de gérer toutes les ressources numériques produites par les établissements universitaires,
- de les partager avec d'autres établissements,
- de les valoriser par une indexation de qualité,
- de les rendre accessibles, à distance et selon les droits définis, dans des interfaces ergonomiques.

En effet, dans les établissements universitaires, l'explosion de la création numérique institutionnelle amène à se poser la question d'une gestion coordonnée et globale de cette production pour en assurer l'accès, sur la durée.

Plusieurs acteurs des établissements sont impliqués dans la production (enseignants, chercheurs, étudiants, le service commun de documentation, le service TICE[1], …) et plusieurs autres sont impliqués dans la valorisation et la gestion pérenne de ces contenus numériques (service commun de documentation, la direction des systèmes d'information, …).

Le développement des UNT (Universités Numériques Thématiques) et des UNR (Universités Numériques en Région) pose la question du partage de ces ressources, dans des systèmes interopérables, accessibles depuis les ENT[2] des établissements et au sein des UNT et des UNR.

[1] Technologies de l'Information et de la Communication pour l'Education

[2] Espace Numérique de Travail : Portail WEB permettant l'accès à des services en relation avec le système d'information de l'établissement

1.1 Historique

ORI-OAI découle de la réflexion de plusieurs initiatives sur la gestion des documents numériques :

- SYNAPSE de l'INSA de Lyon,
- inJAC du consortium ESUP-Portail [2],
- portail de ressources pédagogiques de l'UNT UNIT.

Début 2006, les établissements de ces différents projets ont mis en commun leurs besoins pour développer ORI-OAI.

Ce projet dispose de financements publics issus, notamment, de la sous-direction des bibliothèques de la direction de l'enseignement supérieur et de la sous-direction des TICE de la direction technique du Ministère de l'enseignement supérieur et de la recherche français.

2 Contour fonctionnel

2.1 Les acteurs

Le but du système ORI-OAI est de permettre aux utilisateurs de rechercher des ressources numériques en provenance d'un établissement, ou d'un ensemble d'établissements. Il permet ensuite l'accès à ces ressources de façon libre ou contrôlée.

Au-delà des acteurs qui accèdent en consultation aux ressources, le système est prévu pour permettre à un grand nombre de personnes de publier des ressources. Il peut s'agir d'enseignants, de chercheurs, de personnels administratifs ou même potentiellement d'étudiants.

De plus, la saisie des informations d'indexation plus complexes peut être laissée à un service spécialisé comme, par exemple, les services communs de documentation.

D'une façon générale le système offre une grande flexibilité de configuration afin de laisser le choix à l'établissement de son organisation quant aux processus de publication, de référencement et d'indexation des ressources.

2.2 Les ressources numériques

Le système ORI-OAI a été, depuis le début de sa conception, pensé pour travailler sur différents types de ressources numériques. Chaque ressource pourra, suivant son type, être référencée en utilisant un format de métadonnées adapté.

Les ressources peuvent, par exemple, être des ressources pédagogiques (polycopiés, énoncés et corrigés d'exercices, etc.), des travaux d'étudiants (rapports de stage ou de projets, synthèses bibliographiques, etc.), des travaux de recherche (publications, rapports techniques, mémoires de master, mémoires de thèses, etc.) ou des ressources documentaires éditoriales acquises par l'établissement (périodiques électroniques, livres électroniques, etc.).

2.3 Les fonctionnalités

Le système ORI-OAI est développé afin de répondre à de nombreux objectifs parmi lesquels on peut citer les éléments clés suivants :

Référentiel unique, pour les ressources numériques de l'établissement. Le référentiel ne se substitue pas aux diverses plateformes qui peuvent utiliser et publier ces mêmes ressources (plateformes pédagogiques, sites web des laboratoires …), mais il gère la forme canonique de la ressource et permet éventuellement un contrôle d'accès au sein d'une fédération d'identités,

Système de recherche avancé, multicritères,

Accès thématiques aux ressources, selon des classifications simplifiées qui exploitent la classification Dewey [3],

Système de gestion et de publication des ressources numériques par une publication web avec gestion des droits d'accès ; description des ressources selon les normes LOM [4], TEF [5], Dublin Core [6], etc. en relation avec les autres systèmes documentaires pour le partage des tables d'autorité (SUDOC [7], STAR [8] par exemple) ; Indexation selon les classifications en usage dans les bibliothèques universitaires (Dewey par exemple) et exploitées par les classifications spécifiques des UNT ; archivage des ressources numériques,

Système de production impliquant les acteurs concernés, dans des procédures élaborées, matérialisées par des *workflows* ; gestion des versions et accès aux versions natives des documents pour leurs créateurs,

Système de partage, fondé sur l'échange de métadonnées selon le protocole OAI-PMH [9], permettant de fonctionner au sein d'une communauté constituée (UNT par exemple) en réseau de portails,

Système open-source, libre documenté et pouvant être installé simplement.

3 Concepts

3.1 Le concept de métadonnées

Les métadonnées représentent l'ensemble des informations associées à une ressource. Cet ensemble permet de définir pour chaque ressource un contexte d'utilisation, et un jeu de relations entre les ressources. L'exemple simple est celui du nom d'un auteur, qui permet d'appréhender l'ensemble des ressources qu'il a produites, ou encore celui d'un code issu d'une taxonomie, qui offre une vue sémantiquement structurée d'un ensemble de ressources.

Le concept de métadonnée définit, à travers des formats concrets forgés au fil des pratiques (Dublin Core, LOM, CDM [10], TEF [11], MARC [12], ETDMS [13], etc.), un regroupement déterminé de données structurées, notamment grâce au langage de balise XML. Ce regroupement agrège et anticipe les besoins d'une communauté autour de l'ensemble de ses activités, et assure une sémantique forte grâce à l'utilisation de vocabulaires fermés, standardisés, d'où il tire une très grande efficacité grâce à la qualité de l'information ainsi produite. Ces vocabulaires fermés sont définis à l'intérieur des schémas XML qui valident les métadonnées pour un format donné.

Maintenant, qu'entend-on par entrepôt de données ? C'est l'ensemble des documents, des fiches de métadonnées les décrivant et le système capable d'associer ces documents à ces fiches de métadonnées. Il permet de constituer un référentiel de données.

Ainsi, le format Dublin Core non qualifié définit un jeu de 15 métadonnées à vocation très généraliste, facultatives et répétables, qui peut suffire à représenter des relations simples mais non ambiguës dans un ensemble de ressources : un cahier des charges peut communément être décrit par son auteur, ses dates de création et de publication, un code le situant dans la taxonomie des types de documents utilisés dans l'établissement où il est produit, et quelques autres métadonnées.

Plus complexe, le format LOM est le fruit de la conceptualisation de toutes les informations pertinentes pour décrire un objet pédagogique ; sa structure riche, détaille chacun des aspects par lesquels cet objet peut être appréhendé dans l'ensemble des activités dont il relève : aspect pédagogique pour la gestion scolaire et les conditions pédagogiques d'utilisation (niveau, type de destinataire, durée) qui en détermine l'auditoire, aspect technique pour ses conditions matérielles d'utilisation, aspect juridique pour les conditions légales, aspect du cycle de vie pour ses évolutions, ses versions… Ce format va jusqu'à définir un aspect de méta-métadonnées, pour décrire des informations sur les métadonnées elles-mêmes.

Pour résumer, les métadonnées forment donc la base d'un référencement de qualité, en décrivant de façon adaptée chaque ressource, lui donnant ainsi sa place dans un réseau de relations au sein d'une activité, ce qui permet aux acteurs de cette activité, producteurs comme consommateurs, de retrouver et de manipuler chaque ressource le plus aisément possible.

3.2 Le protocole OAI-PMH

Le protocole OAI-PMH [9] a été créé par l'organisation Open Archive Initiative [14], dédiée aux archives ouvertes et à la mise en œuvre pratique du partage de ressources. Chaque ressource est à cet effet représentée par un **enregistrement** OAI, composé d'un entête OAI et d'un jeu de métadonnées comme précédemment décrit. L'entête contient des informations utilisées par le protocole OAI pour les critères de sélection tels que date de modification, ensembles auxquels l'enregistrement appartient, ou encore son statut supprimé. Chaque format de métadonnées doit être décrit par un schéma XSD dont l'espace de nom et l'adresse internet du fichier sont définis dans les réponses OAI, ce qui rend éligible comme format OAI n'importe quel format décrit par un tel schéma.

A l'aide d'une syntaxe composée de six verbes, simples mais efficaces, OAI-PMH permet de disséminer ces enregistrements par un mécanisme faisant intervenir un fournisseur de service, le **moissonneur**, et un ou plusieurs **entrepôts**, chacun associé à un fournisseur de données.

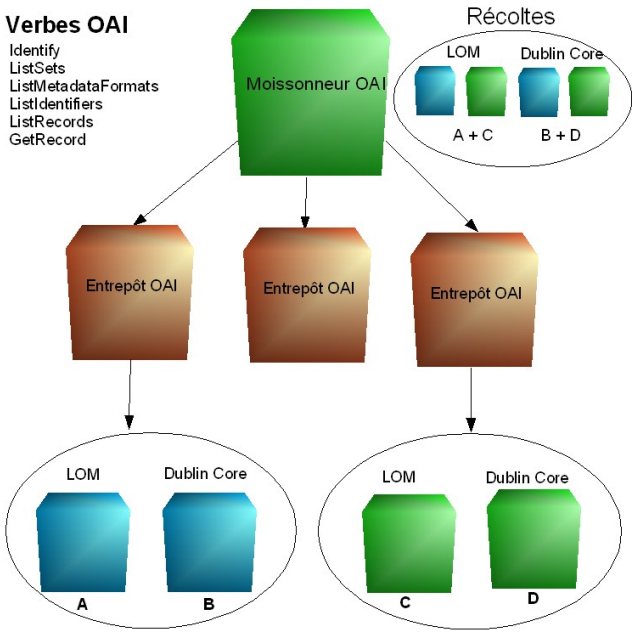

Figure 1 – *Schéma OAI*

Comme le montre la Figure 1, le mécanisme général fait donc intervenir un moissonneur OAI qui récolte de façon sélective des fiches stockées dans plusieurs entrepôts OAI, selon différents **critères de sélection**. Les critères de sélection d'une moisson doivent définir au minimum le **type de format** de métadonnées désiré (comme par exemple ici LOM et Dublin Core), et optionnellement un **ensemble** qui cible plus précisément un domaine ou toute forme de regroupement pertinent pour une utilisation particulière, ainsi que des **dates** de début et de fin. La notion d'ensemble (set OAI) relève d'une fonctionnalité assumée de façon interne par l'entrepôt, pour regrouper ensemble des enregistrements dont le format de métadonnées commun possède des éléments permettant une classification.

Les critères de date permettent d'opérer des moissons incrémentales, c'est-à-dire de ne pas récupérer à chaque fois tous les enregistrements, mais seulement ceux qui ont été ajoutés, modifiés ou supprimés depuis la dernière moisson, ceci afin que le moissonneur puisse traiter correctement le cycle de vie des enregistrements récoltés.

Le premier verbe est ***Identify***, qui fournit une fiche d'identité de l'entrepôt, pouvant comporter une description des contenus disponibles ou des liens sur d'autres entrepôts « amis ». Deux verbes, ***ListSets*** et ***ListMetadataFormats***, fournissent respectivement une liste des ensembles disponibles et une liste des formats supportés. L'accès aux fiches se fait par les verbes ***ListRecords*** et ***ListIdentifiers***, le deuxième ne renvoyant que l'entête muni de l'identifiant de l'enregistrement. Grâce à cet identifiant, on peut récupérer une seule fiche individuellement avec le sixième verbe ***GetRecord***.

Les verbes OAI qui retournent des listes utilisent un jeton de continuation (*resumption token*) pour découper les listes trop longue en plusieurs morceaux, plus facilement manipulables dans les flux sur le réseau.

3.3 Fédération d'identités avec Shibboleth

L'objectif d'une fédération d'identités est de faciliter le partage de ressources numériques en ligne entre établissements en interconnectant leurs services d'authentification. Il devient possible d'ouvrir l'accès à une ressource (pédagogique, scientifique, éditoriale, application métier, etc.) à une population identifiée, sans devoir gérer localement l'enregistrement des utilisateurs.

La fédération d'identités concrétise, pour un groupement d'établissements, l'interconnexion de leurs services d'authentification et l'utilisation d'un ensemble commun d'attributs utilisateurs. Un établissement qui gère un ensemble d'utilisateurs est appelé fournisseur d'identités. Un fournisseur de services est une entité - établissement, administration, société privée - qui propose une ressource numérique en ligne au sein de la fédération. Techniquement, les relations de confiance entre les membres d'une fédération reposent sur des certificats électroniques et une définition partagée des différents fournisseurs d'identités et de services. En outre, la confiance s'établit entre les participants de la fédération au travers d'une convention.

Un même établissement peut participer à plusieurs fédérations et gérer des partenariats de manière bilatérale. Il peut également jouer à la fois le rôle de fournisseur d'identités et de fournisseur de services.

En France, pour l'enseignement supérieur, les règles d'usage et l'assistance technique sur la fédération d'identités sont fournies par le CRU [15]. L'implémentation technique choisie par le CRU comme par le projet ORI-OAI est celle développée dans le contexte universitaire de Internet2, à savoir : Shibboleth [16].

Le schéma suivant décrit de façon très simplifiée le mécanisme de fédération d'identités :

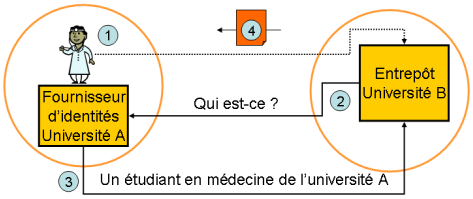

Figure 2 – *Le mécanisme Shibboleth*

1. L'utilisateur tente d'accéder au document dans un autre établissement
2. L'établissement B demande des informations sur l'émetteur de la requête au fournisseur d'identités de l'établissement d'appartenance de l'utilisateur
3. L'établissement A fournit des attributs sur l'utilisateur courant au fournisseur de services de l'établissement B
4. Le fournisseur de services de l'établissement B vérifie ces attributs afin de renvoyer le document ou pas.

3.4 Exemples de contextes de mise en application

Un moissonneur permet de rassembler, en un lieu unique, un ensemble de métadonnées provenant de différents entrepôts. Il est alors possible d'indexer et de proposer une interface de recherche sur ces métadonnées afin de localiser et donner accès à la ressource afférente.

Il est possible d'imaginer de nombreuses mises en application. Nous allons en décrire deux à titre d'exemple :

- Un internaute consulte le site web d'un consortium (exemple : regroupement d'universités sur une thématique ou sur une région) à la recherche de ressources à *priori* libres.
- Un utilisateur d'un établissement, dans le cadre des services proposés par son ENT, recherche des ressources dans son propre établissement et les établissements partenaires. Dans ce cas, les ressources peuvent être soumises à un contrôle d'accès.

3.4.1 Site Web d'un consortium

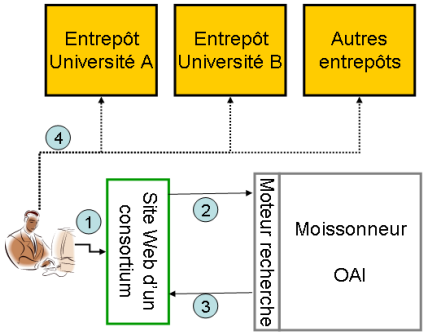

Figure 3 – *Le contexte d'un site Web*

1. L'utilisateur accède au site web du consortium,
2. Il formule une recherche sur le moteur de recherche du consortium,
3. Il obtient des références vers des ressources,
4. Il accède à ces ressources dans les différents entrepôts présents dans les établissements membres du consortium.

3.4.2 Accès depuis un ENT d'établissement

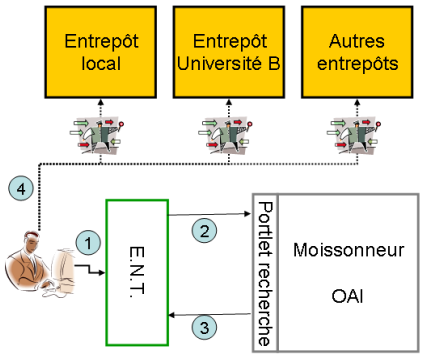

Figure 4 – *Le contexte d'un ENT*

1. L'utilisateur s'identifie pour accéder à son portail de services,
2. Il formule une recherche sur le *portlet* (JSR 168) [17] de recherche ORI-OAI installé dans son établissement,
3. Il obtient des références vers des ressources,
4. Il accède à ces ressources dans l'entrepôt de son établissement ou des établissements partenaires. Si la ressource accédée nécessite un contrôle d'accès les mécanismes de la fédération d'identités sont utilisés de façon transparente pour l'utilisateur (il n'a pas besoin de s'identifier à nouveau).

4 Les enjeux des échanges avec OAI-PMH

Pourquoi avoir choisi le protocole OAI-PMH pour l'échange de données dans ORI-OAI ?

- Assurer la gestion par chacun des données qu'il produit (subsidiarité)

Les établissements en tant que producteur de données gèrent des documents et des fiches de métadonnées qui viennent décrire et qualifier ces documents. Le protocole OAI-PMH s'attache à diffuser ces fiches de métadonnées, indépendamment de la diffusion des documents auxquels ces fiches font référence. L'établissement producteur reste seul gestionnaire (stockage, droits d'accès, …) des documents dont il souhaite diffuser la description. Les établissements restent également maître des fiches de métadonnées diffusées qu'ils ont produites, mais elles sont communiquées à la communauté. Le protocole OAI-PMH assure le suivi des créations, modifications, suppressions de fiches pouvant survenir dans l'entrepôt de données. Cette distinction de gestion entre le document et sa fiche descriptive permet d'éviter de créer des doublons/copies des documents sur les différents portails, de gérer la mise à jour de ces documents… Mais laisse cette responsabilité au producteur du document.

Le protocole OAI-PMH permet également à chaque producteur de données de sélectionner les fiches de métadonnées qu'il souhaite diffuser, indépendamment toujours des droits d'accès donnés aux documents eux-mêmes. Et il permet de proposer un découpage de ce fonds de fiches selon les critères qui lui semblent pertinents pour permettre aux établissements moissonneurs de ses fiches de n'agréger qu'une partie de ce fonds.

C'est donc bien dans tous les cas l'établissement producteur qui reste seul maître des documents qu'il produit et des informations qui sont diffusées sur ces documents. Il contrôle son image sur la toile et dans ses réseaux. Le protocole OAI-PMH permet de faire évoluer ses documents et ses métadonnées sans bouleverser les pratiques et les structures établies.

- Faciliter l'agrégation de données issues des réseaux des établissements

Le protocole OAI-PMH assure aux établissements qui souhaitent agréger des données sur des documents d'autres établissements, d'accéder à une description minimale des documents au format Dubin Core non qualifié et la possibilité d'obtenir des descriptions plus complètes à travers d'autres formats de métadonnées, tous structurés en XML. Les formats des données échangées sont standardisés et connus. Plusieurs formats peuvent être proposés pour une même ressource, ceci permettant d'intégrer plus facilement les ressources à sa propre structure de données.

Il permet à travers différents moteurs de recherche spécifiques (tels OAIster) de découvrir les entrepôts existants, de les parcourir et les analyser pour organiser au mieux sa politique de moissonnage. De même, grâce à la simplicité des verbes d'action proposés par le protocole, de nombreuses plateformes ont été créées pour permettre cette analyse (voir l'outil OAI – Repository Explorer http://re.cs.uct.ac.za/). Chaque établissement fait le choix des entrepôts qu'il moissonne et, dans chaque entrepôt, des ressources qu'il moissonne par l'utilisation des ensembles proposés par celui-ci.

L'agrégation de données est facilitée par la répartition des rôles que nous avons vue précédemment, à savoir que chacun reste seul gestionnaire des données qu'il produit et propose à la diffusion.

- Permettre de construire une politique documentaire solide

Le protocole OAI-PMH permet de sélectionner les fiches de métadonnées produites à diffuser, à quel public (internet, réseau d'établissement, partenariat, échanges internes), sous quelle forme (quel format de métadonnées ?) et de les présenter selon des ensembles qui permettront une sélection plus fine pour les établissements agrégateurs. Tout en gardant la totale gestion de ces fiches, de leurs évolutions et des documents qu'elles décrivent.

Le protocole OAI-PMH permet d'identifier des entrepôts de fiches de métadonnées libres d'accès, de les analyser ainsi que les documents qu'elles décrivent, de récupérer ces fiches en tout ou partie pour les intégrer à une plateforme de recherche locale et enrichir ainsi le catalogue proposé aux utilisateurs.

Pour que diffusion et agrégation puissent avoir lieu, il est nécessaire d'homogénéiser les données (éléments de description, vocabulaires utilisés) afin de proposer un fonds cohérent (par exemple pour un type de ressource). Mais le même système de partage est utilisable pour des données hétérogènes. OAI-PMH permet en effet de diffuser indifféremment des fiches de métadonnées en Dublin Core, en LOM ou tout autre format interne ou partagé utilisé.

Le protocole OAI-PMH permet la construction de catalogues à travers la collecte de différentes sources (internes et externes) pour proposer ensuite une base commune d'interrogation et/ou des moteurs de recherche spécifiques.

- Simplicité technique

La mise en place de ce système d'agrégation est simplifiée par sa légèreté technique, la légèreté des données échangées (uniquement les métadonnées sur les documents), la mise à jour asynchrone, planifiée et automatisée. Il peut s'ajouter à tout système de gestion des données existant comme une couche supplémentaire de fonctionnalités.

« C'est un protocole simple et ouvert. Il utilise des technologies ouvertes qui sont des standards reconnus sur le web (protocole HTTP, langage XML). Il offre une grande liberté d'application et une grande simplicité de mise en œuvre.

C'est un protocole largement répandu »

E. Bermès, BnF, OAI Open Archive Initiative, Journée d'information AFNOR CG46 7 juin 2005, http://www.bnf.fr/PAGES/infopro/journeespro/pdf/AFNOR2005/OAI.pdf

5 Mise en œuvre

5.1 L'architecture globale du projet

L'architecture du projet ORI-OAI se présente sous la forme de huit composants développés indépendamment. Chacun de ces composants a un rôle bien défini dans le système et communique avec les autres au travers de services exposés en frontal de chaque composant. La technologie choisie pour rendre possible le dialogue entre tous les modules est les Web services [18].

Ce choix permet une grande souplesse dans d'éventuelles déclinaisons d'architectures et de langages de programmation. En effet, cette architecture technique a été notamment pensée pour permettre à des logiciels extérieurs de dialoguer avec différents éléments du système.

Soulignons que ORI-OAI est développé en Open Source sous licence GPL (General Public License) [19].

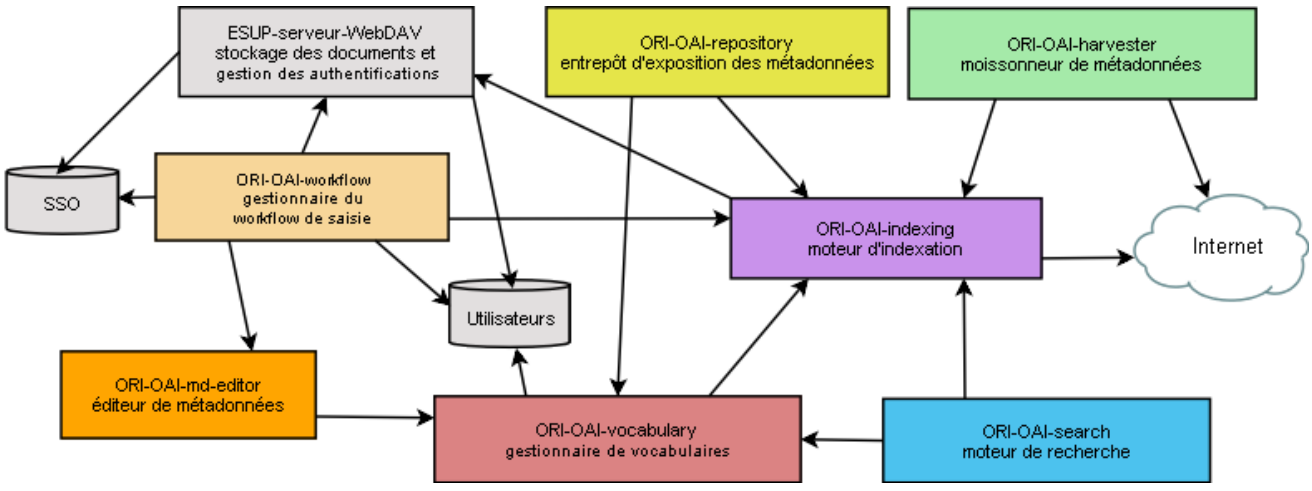

Figure 5 – *L'architecture globale*

5.2 Les différents modules

Nous avons vu que tous les modules sont développés indépendamment les uns des autres, cependant, ils reposent sur certaines technologies communes. En effet, le langage de développement Java [20] ainsi que l'environnement de développement Spring [21] représentent les technologies de base de chaque composant. De plus, toutes les communications par Web services sont implémentées via XFire [22].

Voyons maintenant les fonctionnalités de chaque module ainsi que les technologies qu'ils emploient. Soulignons que toutes ces technologies sont des projets Open Source.

ORI-OAI-workflow

Ce composant est utilisé pour le workflow de référencement des ressources. Les technologies utilisées par ce moteur de publication sont diverses et permettent un paramétrage très fin, complet et puissant en fonction des besoins de l'établissement ou du consortium qui le met en place.

Il sera aussi couplé dans la version 1.5 avec l'outil de gestion des documents dans le système et l'affectation des droits d'accès aux ressources. En attendant cette version, il est possible uniquement de renseigner le lien vers le(s) document(s) depuis l'interface de saisie des métadonnées ; le dépôt devant se faire en dehors du workflow. Dans le cas où vous utilisez le serveur WebDAV pour gérer vos documents, il est possible d'utiliser n'importe quel client compatible avec la norme WebDAV en dehors du workflow pour déposer la ressource et copier le lien vers cette ressource dans la fiche de métadonnées. Pour les établissements utilisant l'ENT ESUP Portail, vous avez la possibilité d'utiliser le canal stockage pour faire le dépôt de la ressource et gérer les droits d'accès (lecture/écriture pour des usagers, groupes d'usagers de votre établissement ou externes en utilisant la fédération d'identités avec shibboleth).

Le moteur de *workflow OsWorkflow* [23] utilisé dans ce module permet une configuration avancée de toutes les actions effectuées à chacune des étapes de la publication. En effet, par de la configuration, il peut répondre aux besoins des établissements exigeant du déposant la saisie de tous les champs de métadonnées, mais aussi des établissements qui souhaitent une gestion du dépôt plus structurée où la saisie des métadonnées s'effectue en différentes étapes, par différents intervenants. Le formulaire de saisie des métadonnées est supporté par un appel à ORI-OAI-md-editor aux différentes étapes du workflow. Chaque utilisateur aura alors, pour une même fiche, une vue différente. Ceci lui permettra de voir et d'éditer uniquement les métadonnées sur lesquelles il a un droit d'accès particulier.

Les autres technologies utilisées dans ce module sont Acegi Security [26] pour gérer toutes les sécurités d'accès à l'application, une base de données SQL pour le stockage des fiches de métadonnées et les données de gestion, Hibernate [28] pour rendre les accès à la base de données relationnelle transparents, Compass/Lucène pour réaliser des requêtes rapides et efficaces sur les fiches de métadonnées associées aux données de gestion et enfin JSF [29] pour la gestion des interfaces graphiques.

ORI-OAI-md-editor

Ce composant est utilisé pour la saisie des métadonnées dans ORI-OAI. Il propose des formulaires de saisie de métadonnées entièrement configurables. La technologie XForms [24] associée au moteur de transformation Orbeon [25] offre des formulaires de saisie dynamiques en fonction de fichiers XML.

Le support des différents formats de métadonnées repose sur la possibilité d'enrichir le module ORI-OAI-md-editor de nouvelles configurations XForms.

Ce module permet une interface de saisie des métadonnées très riche grâce notamment à de l'aide à la saisie, de l'auto complétion, de la recherche de personnes dans l'annuaire, etc.

Couplé au module ORI-OAI-workflow, il permet de proposer un formulaire de saisie des métadonnées à n'importe quelle étape du workflow. Utilisé seul, il permet l'édition de fiches de métadonnées en dehors du système. Il répond alors au besoin d'un éditeur simple et puissant de métadonnées au format XML.

ORI-OAI-indexing

Une fois le dépôt de ressources et la saisie de métadonnées validés, ces dernières sont indexées par le module ORI-OAI-indexing. Ce module a pour rôle l'indexation des fiches de métadonnées ainsi que des documents associés.

Pour cela, il utilise le moteur d'indexation Lucene [30]. Celui-ci permet l'indexation de différentes sources offrant une recherche puissante et rapide en se reposant sur différents analyseurs. L'analyseur de la langue française permettra notamment la gestion des verbes conjugués, des pluriels ou encore des accents et caractères spéciaux. Un système de pondération permet aussi de rendre une métadonnée plus pertinente qu'une autre. Par exemple, on préfèrera retrouver en premier les documents dont l'élément recherché se trouve dans le titre plutôt que dans la description.

Le module utilise également Lius [31] qui est un framework d'indexation basé sur le projet Lucene. Il permet une indexation de différents formats de fichiers comme XML, PDF, OpenOffice, ZIP, MP3, etc. Il est utilisé dans notre projet pour offrir une configuration avancée des champs à indexer dans les différents formats de fiches de métadonnées XML et, par la suite, pour indexer les documents associés en plein texte.

En plus de l'aspect indexation, ORI-OAI-indexing offre un service de recherche de documents via Web service en se reposant sur la syntaxe des requêtes Lucene. Il est utilisé par différents composants dans le système.

ORI-OAI-repository

Nous avons vu dans un chapitre précédent le protocole OAI-PMH. Le module ORI-OAI-repository se charge, via ce protocole, de l'exposition des fiches de métadonnées saisies dans le module ORI-OAI-workflow et/ou de celles provenant de moissons OAI. Utilisant le logiciel OAICat [32], il expose les fiches dans le but d'être moissonnés par tout moissonneur OAI.

Ce module gère également le concept de « sets OAI-PMH ». Cet aspect du protocole OAI permet d'exposer les fiches de métadonnées sous forme d'ensembles distincts. Ces ensembles sont souvent liés à une thématique particulière. Nous pouvons par exemple identifier l'ensemble de toutes les fiches pédagogiques au format LOM associées aux notions de mathématiques. Pour identifier les fiches correspondant aux différents ensembles, le module ORI-OAI-repository construit des requêtes suivant les critères associés aux différents sets et les envoie au module ORI-OAI-indexing.

ORI-OAI-harvester

Ce composant du système correspond au moissonneur OAI-PMH. Utilisant le logiciel OAIHarvester2 [33], il permet le moissonnage de fiches de métadonnées sur tout entrepôt OAI. Les fiches moissonnées sont alors stockées localement dans une base de données SQL.

Tout comme le moteur de *workflow*, ce module fournit toutes les fiches de métadonnées moissonnées au moteur d'indexation dans le but d'être indexées.

Via une interface graphique conviviale, l'administrateur du système peut programmer les différentes moissons qui seront lancées par le gestionnaire de tâches Quartz [34].

ORI-OAI-search

Ce module offre une interface graphique pour la recherche de documents dans le système. Dialoguant avec le module ORI-OAI-indexing, il génère des requêtes dans la syntaxe Lucene et affiche les documents retrouvés.

Ce composant est entièrement configurable en ce qui concerne les formats de documents que l'on souhaite rechercher, et les types de recherches que l'on veut proposer à l'utilisateur. Il existe trois types de recherche :

- Par date : on propose à l'utilisateur de rechercher des documents suivant leur date de création, modification, etc. Ce type est utilisé pour afficher les nouveautés.
- Avancée : il est possible de configurer différents formulaires de recherche avec des critères plus ou moins avancés. On pourra proposer par exemple un formulaire de recherche composé d'un seul champ permettant une recherche sur le document complet et les métadonnées associées, ou encore un formulaire de recherche avancée proposant des champs de recherche pour chacune des métadonnées d'un format de description.
- Thématique : dans ce type de recherche, on ne demande aucune saisie à l'utilisateur. Elle est mise en place pour faire des recherches suivant des classifications de documents, des auteurs, des mots-clefs, etc. Elle peut être sollicitée par les utilisateurs souhaitant découvrir les documents référencés dans le système n'ayant aucun critère de recherche particulier. Par ce type de recherche, on guide également l'utilisateur dans ses recherches en ne proposant que les valeurs réellement indexées comme par exemple dans le cas de la recherche par mots-clefs.

Ce module permet donc une recherche multicritères et l'export des résultats de recherche sous forme de flux RSS, dans un catalogue au format RTF ou encore l'export des fiches de métadonnées en XML.

Notons également que le module peut être décliné en deux versions : *servlet* pour une installation standard sur un serveur Web et *portlet* pour une intégration dans un Environnement Numérique de Travail.

ORI-OAI-vocabulary

Le composant ORI-OAI-vocabulary est celui qui gère les vocabulaires utilisés les différents modules. On entend par vocabulaire un ensemble de valeurs disponibles pour un critère donné.

Les vocabulaires reposent sur le format VDEX [27]. Ils peuvent être statiques, ils sont alors configurés via des fichiers XML. C'est le cas des classifications ou des vocabulaires définis par les normes comme le LOM. Ils peuvent aussi être dynamiques. Ils peuvent par exemple s'appuyer sur le contenu d'un index pour proposer les différentes valeurs actuellement indexées. Par exemple, on pourra constituer dynamiquement, en interrogeant le module ORI-OAI-indexing, la liste des mots-clefs libres ou des auteurs qui ont déjà été saisis dans les documents pédagogiques.

Ces vocabulaires sont utilisés par :

- Le module ORI-OAI-md-editor pour proposer des listes de valeurs lors de la saisie des métadonnées.
- Le moteur de recherche ORI-OAI-search pour les recherches thématiques ou les valeurs disponibles pour certains champs de la recherche avancée.
- L'entrepôt ORI-OAI-repository pour générer dynamiquement des sets OAI en fonction par exemple d'une thématique donnée.

ESUP-serveur-WebDAV

Initialement développé par le consortium ESUP Portail, ce serveur permet le stockage des documents en ligne en utilisant le protocole WebDAV [35]. Le socle de ce composant est le serveur Jakarta/Slide [36]. Via différentes méthodes du protocole WebDAV, il est possible de gérer des documents distants (dépôt, téléchargement, suppression, etc.). De plus, le protocole ACP (Access Control Protocol) offre une gestion dynamique des droits (lecture, écriture, etc.) sur les documents.

Les développements qui ont été faits autour de ce serveur sont :

- Le support d'authentifications via le SSO *(Single Sign-On)* CAS [37].
- La gestion de fédération d'identités par Shibboleth.
- L'externalisation de la gestion des groupes d'utilisateurs.
- Le support des quotas.

6 La distribution de ORI-OAI

6.1 Les versions disponibles

La première version 1.0 de ORI-OAI a été rendue publique depuis octobre 2007, suivie par la version 1.1 en juin 2008.

Cette version offre les fonctionnalités d'un outil de référencement pour des ressources pédagogiques, ainsi qu'à tout type de ressource documentaire, grâce aux jeux de métadonnées du standard LOM et du Dublin Core. Avec une implémentation complète du protocole OAI PMH, ainsi que l'utilisation des Web Services, il pose les bases d'une large interopérabilité avec d'autres systèmes. Son système de gestion de *workflow* permet de configurer de façon très souple les rôles des différents intervenants lors de la production et lors de la validation des métadonnées, dont la qualité est encore renforcée par l'utilisation de vocabulaires standardisés et partagés par l'ensemble du système.

6.2 Les évolutions futures

La version 1.5 à venir va proposer une refonte complète du système de stockage en utilisant la plate-forme Nuxeo [40]. Ce système sera couplé au gestionnaire de workflow pour une intégration complète du dépôt des documents et de leur référencement. Aussi, un connecteur sera proposé pour permettre à une application extérieure (comme par exemple la plateforme pédagogique Moodle [39]) de démarrer un processus de référencement dans le workflow.

Dans la version 1.5, le moteur d'indexation proposera une indexation plein texte des documents. Le moteur de recherche proposera quant à lui d'autres possibilités de recherche.

L'équipe ORI-OAI travaille aussi sur la gestion d'autres formats de métadonnées pour le référencement des thèses, des publications de la recherche ou encore des documents administratifs. La future version de ORI-OAI proposera aussi l'interopérabilité avec des systèmes nationaux comme HAL [38] pour les publications ou STAR pour les thèses.

6.3 Les utilisateurs

A ce jour, la plate-forme ORI-OAI est utilisée ou en cours de mise en place par une vingtaine d'établissements de l'enseignement supérieur français pour le référencement des ressources pédagogiques. Quatre UNT et deux UNR utilisent également ORI-OAI.

L'Université Virtuelle de Tunis et le Campus Virtuel Marocain utilisent également ORI-OAI comme portail de ressources pédagogiques.

Bibliographie

[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.sudoc.abes.fr
[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.abes.fr/abes/documents/tef/index.html
[12] http://www.loc.gov/marc
[13] http://www.ndltd.org/standards/metadata/etd-ms-v1.01.html
[14] http://www.openarchives.org
[15] http://federation.cru.fr
[16] http://shibboleth.internet2.edu/
[17] http://jcp.org/en/jsr/detail?id=168
[18] http://www.w3.org/2002/ws/
[19] http://www.gnu.org/licenses/gpl.html
[20] http://www.java.com/
[21] http://www.springframework.org/
[22] http://xfire.codehaus.org/
[23] http://www.opensymphony.com/osworkflow
[24] http://www.w3.org/MarkUp/Forms/
[25] http://www.orbeon.com/
[26] http://www.acegisecurity.org/
[27] http://www.imsglobal.org/vdex/
[28] http://www.hibernate.org/
[29] http://java.sun.com/javaee/javaserverfaces/
[30] http://lucene.apache.org/java/docs/index.html
[31] http://www.bibl.ulaval.ca/lius/
[32] http://www.oclc.org/research/software/oai/cat.htm
[33] http://www.oclc.org/research/software/oai/harvester2.htm
[34] http://www.opensymphony.com/quartz/
[35] http://www.webdav.org/
[36] http://jakarta.apache.org/slide/
[37] http://www.ja-sig.org/products/cas/
[38] http://hal.archives-ouvertes.fr
[39] http://moodle.org
[40] http://www.nuxeo.com