
	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01

Étude format d'échange compatible OAI-PMH


Livrable du au titre du projet	COCLICO
Sous-projet	WP2
Tache	2.6
Livrable	L2.6.1

Rédacteur(s)	Vérificateur(s)	Approbateur(s)
Luca Dini Hratchia Pélibossian Sebastien Dabet Celi-France	O.Berger	

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01


Documents applicables
Annexe technique au projet COCLICO

Documents de références (pour information)
http://lucene.apache.org/solr/
http://lucene.apache.org/solr/tutorial.html
http://www.openarchives.org/
http://www.openarchives.org/OAI/openarchivesprotocol.html
http://www.dublincore.org/
https://forge.projet-coclico.org/projects/wp3/
http://www.projet-coclico.org/index.php/DESCRIPTION_TECHNIQUE_DU_PROJET
https://forge.projet-coclico.org/projects/wp2/

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01


Gestion des versions

N° de version	Date	Auteurs	Modification apportées
0.01	20/05/10	Pélibossian	Première version

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01

Sommaire

Objectif	5
Contexte	5
1. Fonctionnalités pour la fédération des forges	6
1.1. OAI-PMH	7
1.2. Index de Solr comme un entrepôt OAI pour les métadonnées de Forge	7
2. Format d'échange compatible OAI-PMH	10
2.1. Dublin Core	11
2.2. Métadonnées de forges et Dublin Core	12
2.3. Quelques outils open-source pour construire portail de fédération des Forge compatible OAI-PMH?	14

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01


Objectif

L'objectif de ce document est de définir un format d'échange des métadonnées de forges simple, efficace, générique et compatible avec OAI-PMH sur lequel nous construirons de fonctionnalités pour la fédération des forges. Les fonctionnalités pour la fédération des forge permettrons :

- Interpréter les meta-données venant des forges.
- Produire des index thématiques, selon la volonté du site fédérateur.
- Permettre la création rapide de portails thématiques qui fédèrent différentes forges et projets.
- Permettre la mise en place d'un service « Mes projets » où tous les projets d'un certain utilisateur ou groupe sont indexés (même s'ils résident sur des forges différentes). »

Contexte

Ce document est le résultat de travaux effectués dans la tâche 6 « Fonctionnalités pour la fédération de forges » du sous-projet 2 « Interopérabilité et échange de données ».

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01

1 Fonctionnalités pour la fédération des forges

Nous pouvons constater que il y a plus en plus des projets qui vise de fédérer les logiciels libre opens sources pour des but divers et varier. Voici quelques exemples:

OHLOH est un répertoire public et gratuit de logiciels open source, des personnes, une communauté, et un service d'analyse. Ils utilisent les données de leur annuaire pour créer des rapports historiques sur la démographie changeante du monde open source.

FSF (*Free Software Foundation*) est une organisation à but non lucratif dédiée aux Logiciels Libres. Leur travail consiste à faire comprendre l'intérêt des Logiciels Libres et à provoquer le soutien de la liberté des logiciels auprès des politiques, dans les lois et dans la société dans son ensemble. FSF ont des catalogues de logiciels libres ou on trouve un certains nombre informations sur divers solutions

ADULCAT est une association des développeurs et utilisateurs de logiciels libres pour l'administration et les collectivités territoriales il propose des atelier-forge basé sur fusion forge pour le développement et la distribution des logiciels.

OSOR (*Open Source Observatory and Repository*) peut être divisé en deux composantes principales:

La Plate-forme d'information, destinées aux administrations publiques fournit des nouvelles, des conseils, liens, contacts, etc et une **bibliothèque** où les logiciels (source et code objet), la documentation et la connaissance est facilement accessible selon une taxonomie spécifique du secteur public.


Nous pouvons également trouver d'autres exemples des fédérations des information sur les logiciels libres et ce n'est pas difficile de prévoir que l'envie de fédérer et de regrouper les informations sur les logiciels libres par rapport aux différentes critères ne va pas manquer et certainement une forte quantité de ces projets seront pratiques et profitables.

En tenant compte que la quasi-totalité des projets Open Source se repose sur les forges, nous avons envisagé dans le cadre du projet COCLICO, de mettre à disposition des forges un composant leur permettant d'exposer l'intégralité de leur contenu dans un format que puisse être « interprété » par tout système de fédération des bibliothèques numériques qui soit compatible avec OAI-PMH. Ce composant permettra:

- Interpréter les métadonnées venant des forges.
- Produire des index thématiques, selon la volonté du site fédérateur.
- Permettre la création rapide de portails thématiques qui fédèrent différentes forges et projets.
- Permettre la mise en place d'un service « Mes projets » où tous les projets d'un certain utilisateur ou groupe sont indexés (même s'ils résident sur des forges différentes). »

Nous proposons un système basé sur l'OAI-PMH (Open Archives Initiative's Protocol for Metadata Harvesting) qui est un protocole d'échange et de transfert de données. OAI-PMH est utilisé au niveau mondial par les acteurs du monde des archives électroniques ouvertes pour échanger des métadonnées..

L'architecture l'OAI-PMH centralise les métadonnées décrivant différentes ressources, en laissant ces ressources à leur emplacement initial par conséquent la publication d'information pour les utilisateurs sera effectué en local, nous pensons que c'est un avantage très appréciable. Il a également une procédure de collecte d'information qui permet mettre en jour ces données avec une fréquence paramétrable.

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01

1.1 OAI-PMH

OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*) est un protocole d'échange et de transfert de données. Il vise à rendre partageable dans le monde entier des descriptions de documents (métadonnées) et à les rendre accessibles, sans les dupliquer ni modifier leur localisation d'origine. Il aide également à la synchronisation des différentes sources de données.

Ce protocole, basé sur les notions d'**entrepôt** et de **moissonneur**, permet d'harmoniser l'accès à des sources hétérogènes de données indépendamment des applications utilisées.

Pour cela, ce protocole définit les conditions de collecte et de transfert de métadonnées sous forme d'enregistrements d'une archive ouverte, produite par un fournisseur de données, vers le serveur d'un fournisseur de services.

Fonctionnellement, le protocole centralise les métadonnées décrivant différentes ressources, en laissant ces ressources à leur emplacement initial. Seules les métadonnées sont rassemblées par le fournisseur de services qui les exploite pour ses besoins spécifiques en donnant accès aux ressources dans un contexte particulier.

L'OAI-PMH définit trois acteurs différents dans une architecture de l'information basée sur les métadonnées :

Le fournisseur de données (ou fournisseur de contenus - data providers)

Le fournisseur de données définit les données exposées, le(s) schéma(s) selon lesquels les données sont mises à disposition, et les rend accessibles sur un **entrepôt OAI** qu'il maintient à jour.

Le fournisseur de service (ou service providers)

Le fournisseur de services lance un programme appelé **moissonneur** pour envoyer une requête à un fournisseur de données et en collecter les métadonnées. Le fournisseur de services traite les métadonnées qu'il a rassemblées et offre un service basé sur ces métadonnées. Par exemple une création d'un portail, un service de création et de mise à jour de catalogues et de bibliothèque des données, un service d'indexation et de recherche fédérée, un service d'accès et de visualisation des données.

L'agrégateur

L'agrégateur est un troisième type d'acteur qui peut, dans certains cas, s'ajouter à cette configuration : il fournit des données intermédiaires. Il rassemble les métadonnées provenant de plusieurs fournisseurs de données et les rend accessibles dans un entrepôt OAI, éventuellement après les avoir retraitées. Les agrégateurs peuvent garantir des métadonnées de meilleure qualité, assurer un stockage commun, engager un travail de normalisation.

Voici un schéma pour un portail de fédération des forges basé sur l'OAI-PMH

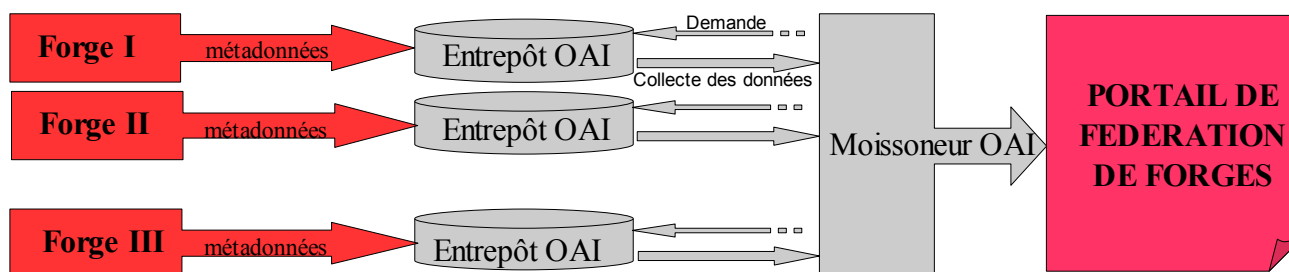



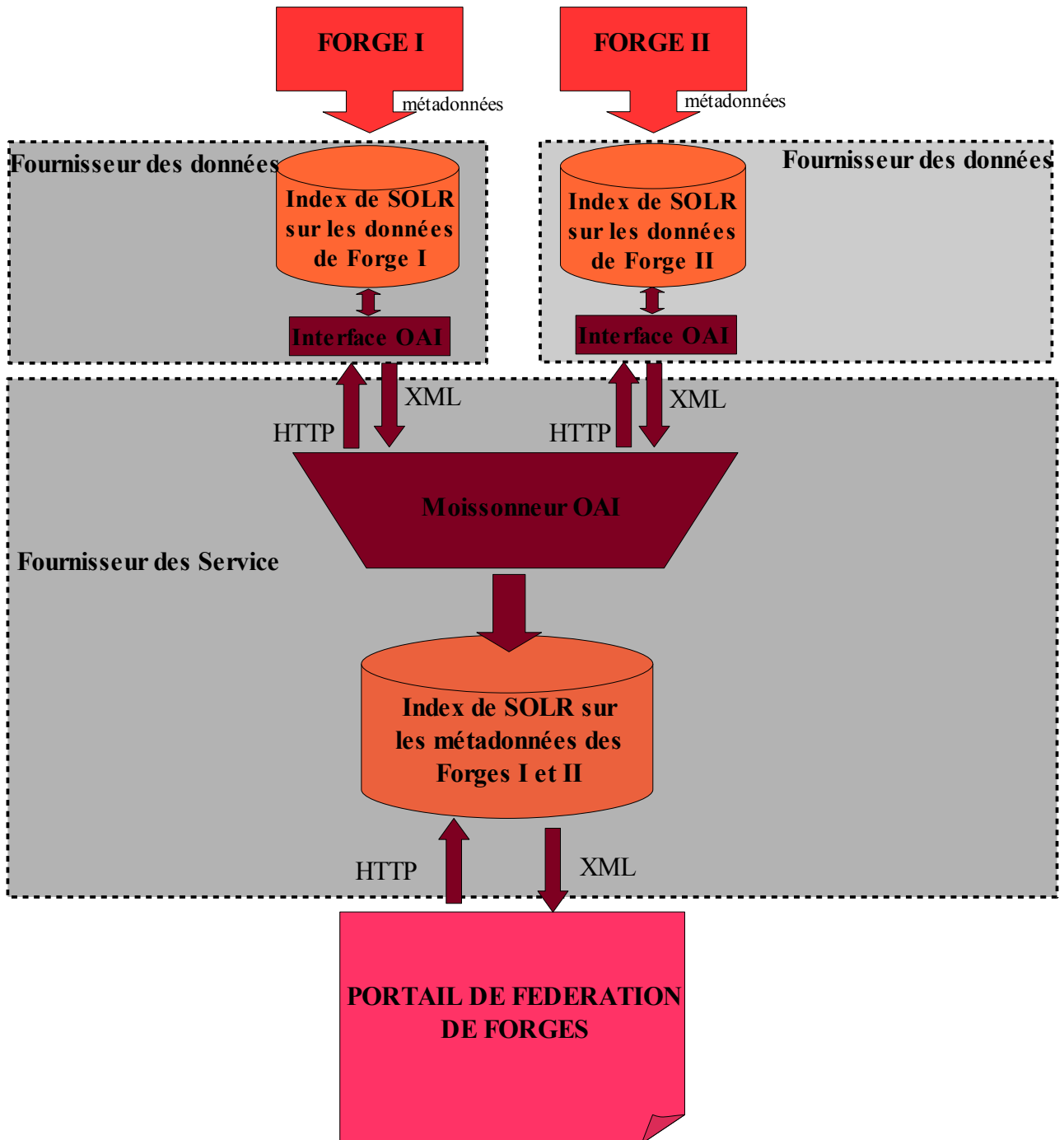
Schéma 1: Architecture OAI-PMH pour un portail de fédération de forges

1.2 Index de Solr comme un entrepôt OAI-PMH pour les métadonnées de Forge

Dans le cadre du WP3, tâche 3.4.1 nous avons adopté Solr comme un moteur de recherche dans le forge et


	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01

nous avons proposé de stocker les métadonnées de forge dans l'index de Solr pour le besoin de recherche d'information locale (livrable 3.4.2). Cela nous permet d'envisager l'index de Solr comme un entrepôt des métadonnées de forge et de construire un entrepôt OAI-PMH sur l'index de Solr.



L'api entre Index de SOLR et le portail de fédération sera l'api SOLR qui se base sur HTTP/XML.


Cette architecture nous permet de considérer l'index de Solr de chaque forge comme un entrepôt de métadonnées et de construire un serveur Solr pour stocker les métadonnées de différentes forges. Ce serveur Solr servira comme fournisseur de service. Le protocole OAI-PMH permet uniquement au fournisseur de

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01

services de mettre à jour son entrepôt local (index de solr) nourri par entrepôt des données de chaque Forge. L'ensemble du protocole OAI-PMH repose sur trois normes :

- Le protocole HTTP qui permet l'envoi des requêtes et un retour de données structurées au format XML.
- Le schéma XML pour l'implémentation et l'échange des métadonnées.
- La norme Dublin Core pour la description des ressources.

Pour la réalisation de cet architecture nous avons besoins de définir un format d'échange de données compatible OAI-PMH. Le format Dublin Core nous parait suffisant pour cette tâche. **Interface OAI** dans le schéma nous servira à transformer les données stockés localement dans l'index de Solr des forges en format Dublin Core qui est compatible avec OAI-PMH. Les spécifications des interfaces OAI sont présenté ici <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01

2 Format d'échange compatible OAI-PMH

Les ressources de forges sont très variées et de ce fait, la description des ressources diffère et varie en fonction du type et de l'usage de la ressource. Les standards concernant les métadonnées sont généralement orientés "métiers", ils s'appliquent à un domaine de connaissances ou d'actions plus ou moins ciblés. En définissant les éléments de description et en recommandant les règles d'utilisation, ces normes de métadonnées fournissent les mécanismes nécessaires au partage de l'information.

Parmi les schémas de métadonnées les plus courants, on peut distinguer Dublin Core comme un schéma de métadonnées permettant de décrire tout type de document numérique.

2.1 Dublin Core

Le Dublin Core est le schéma de métadonnées le plus couramment adopté permettant de décrire une grande variété de ressources numériques. Cette norme a été établie par un consensus international de professionnels œuvrant dans diverses disciplines (bibliothèque, informatique, musée, etc.). L'objectif initial du Dublin Core était de définir un ensemble d'éléments simples et concis permettant aux auteurs de décrire leurs propres ressources électroniques.

Le Dublin Core constitue un format standard facile à implémenter et efficace pour décrire simplement des ressources et d'identifier les responsabilités légales attachées au document.

Le Dublin Core simple comprend 15 éléments pouvant être optionnels, répétés et présentés dans n'importe quel ordre. De plus, chaque élément possède un ensemble d'attributs permettant de raffiner la signification de l'élément. Ces éléments sont repartis en 3 groupes :


- Le contenu comprenant les éléments : *couverture, description, type, relation, source, sujet et titre.*
- La propriété intellectuelle avec les éléments : *collaborateur, créateur, éditeur et droits.*
- L'instance particulière avec les éléments : *date, format, identifiant et langue.*

Exemple de représentation en format RDF+XML:

```
<?xml:namespace href="http://www.w3c.org/RDF/" as="RDF"?>
<?xml:namespace href="http://purl.org/RDF/DC/" as="DC"?>
<RDF:RDF>
  <RDF:Description>
    <dc>Title>Dublin Core Metadata Element Set: Reference Description</dc>Title>
    <dc:Creator>Stuart Weibel</dc:Creator>
    <dc:Creator>Eric Miller</dc:Creator>
    <dc:Subject>Metadata, Dublin Core element, resource description</dc:Subject>
    <dc>Description>This document is the reference description of the
      Dublin Core Metadata Element Set designed to facilitate
      resource discovery.</dc>Description>
    <dc:Publisher>OCLC Online Computer Library Center, Inc.</dc:Publisher>
    <dc:Identifier>http://purl.org/metadata/dublin_core_elements</dc:Identifier>
    <dc:Format>text/html</dc:Format>
    <dc>Type>Technical Report</dc>Type>
    <dc:Language>en</dc:Language>
    <dc>Date>1997-11-02</dc>Date>
  </RDF:Description>
</RDF:RDF>
```

Grâce à sa simplicité, le Dublin Core est très utilisé et permet de décrire tout type de ressource.

En juillet 2000, l'initiative de métadonnées du Dublin Core (IMDC / DCMI : Dublin Core® Metadata Initiative) a émis sa liste de qualificatifs Dublin Core (Dublin Core qualifié) recommandés. Actuellement l'IMDC reconnaît

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01


deux grandes catégories de qualificatifs:

- **Le raffinement d'éléments.** Ces qualificatifs permettent de préciser le sens d'un élément pour qu'il soit plus circonscrit ou plus précis. Un élément raffiné partage le même sens que l'élément non qualifié mais avec une portée plus restreinte. Un client qui ne peut interpréter le terme raffinant un élément spécifique devrait être capable d'ignorer le qualificatif et de traiter la valeur de la métadonnée comme s'il s'agissait d'un élément non qualifié (plus large). Les définitions des termes de raffinement d'éléments pour les qualificatifs doivent être publiquement disponibles.
- **Le schéma d'encodage.** Ces qualificatifs identifient des schémas qui aident à l'interprétation de la valeur d'un élément. Ces schémas comprennent des vocabulaires contrôlés et des notations formelles ou des règles d'interprétation. Une valeur exprimée en utilisant un schéma d'encodage pourra donc être une expression sélectionnée à partir d'un vocabulaire contrôlé (e.g. un terme d'un système de classification ou un ensemble de vedettes matières) ou une chaîne de caractères formatée en accord avec une notation formelle (e.g. "2000-01-01" comme expression normalisée d'une date). Si un schéma d'encodage ne peut être interprété par un client ou un agent, la valeur peut toujours être utile à un lecteur humain. La description de référence d'un schéma d'encodage utilisé avec des qualificatifs doit être clairement identifiée et disponible pour un usage public.

Voici un table avec les raffinements introduits par l'extension du Dublin Core:

Qualifié	Qualifiants		Qualifié	Qualifiants		
<i>title</i>	alternative	autre forme de titre	<i>coverage</i>	spatial temporal	couverture spatiale couverture temporelle	
<i>description</i>	tableOfContents	table des matières	<i>audience</i>	mediator educationLevel	public visé par la ressource	
	abstract	résumé				
<i>rights</i>	accessRights	restrictions de l'accès à la ressource	<i>date</i>	created	date de création	
<i>relation</i>	isVersionOf	est une version de (édition, adaptation, traduction): changement dans le contenu		valid	date de validité	
	hasVersion	a d'autres versions		available	date de disponibilité	
	isReplacedBy	est remplacé par		issued	date de parution	
	replaces	remplace		modified	date de modification	
	isRequiredBy	est requis par		dateAccepted	date d'acceptation	
	requires	requiert		dateCopyrighted	date du copyright	
	isPartOf	est une partie de		dateSubmitted	date de soumission	
	hasPart	a comme partie		<i>identifier</i>	bibliographicCitation	référence bibliographique de la ressource
	isReferencedBy	est référencé par		<i>format</i>	extent medium	étendue de la ressource (taille, durée) support
	references	référence				
isFormatOf	est un autre format de (changement sur le format par sur le contenu)					
hasFormat	a pour autre format					
conformsTo	est conforme à					

Un des intérêts de l'utilisation du Dublin Core est que sa définition sémantique peut être donné en utilisant le langage de représentation des connaissances RDFS (ressource Description Framework Schema). Il sera donc possible de profiter du mécanisme de spécialisation de RDFS pour déclarer des éléments qui pourront représenter les spécificités des forges de COCLICO.

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01

2.2 Métadonnées de forges et Dublin Core

Les métadonnées des forges qui adopte notre plugin d'indexation sont définies dans le schema.xml de Solr. Nous pouvons les diviser en huit groupes principales:


1. **Forge Générale:**
id, forge_project_id, forge_project_name, document_url, forge_permission_group, forge_permission_group_name, file_name_extention, comments
2. **Document:**
file_address, recordtype, document_type, doc_title, doc_description, doc_owner, doc_owner_name, doc_create_date, doc_update_date, doc_language, doc_extra_field
3. **Mail**
mail_subject, mail_sender, mail_create_date
4. **Forum**
forum_name, forum_subject, forum_body, forum_author, forum_date
5. **News**
news_name, news_subject, news_body, news_author, news_date
6. **Source**
source_file_name, source_owner, source_create_date, source_update_date, source_last_committer,
7. **Release**
release_package_name, release_release_name, release_release_name_address, release_owner, release_date
8. **Tracker**
tracker_title, tracker_identifiant, tracker_type, tracker_description, tracker_subject, tracker_creator, tracker_created, tracker_modified, tracker_name, tracker_project, tracker_component, tracker_status, tracker_owner, tracker_priority, tracker_severity, tracker_attachments_address, tracker_relatedChangeRequests, tracker_changeSets, tracker_comments,

D'après ce classement de métadonnées et compte tenu du classement de Dublin Core suivant:

- **Le contenu:** *titre, sujet, description, source, type, relation, couverture.*
- **La propriété intellectuelle:** *créateur, collaborateur, éditeur et droits.*
- **L'instance particulière:** *date, format, identifiant et langue.*

nous pouvons créer une correspondance (mapping) entre les métadonnées de forge et des éléments de Dublin Core:

- ✓ **dc:identifiant:**
document_url
- ✓ **dc:title:**
doc_title, forum_name, news_name, source_file_name, release_package_name, tracker_title
- ✓ **dc:subject:**
forum_subject, news_subject, tracker_subject
- ✓ **dc:description:**
doc_description, forum_subject, news_subject, source_file_name, release_release_name, tracker_description
- ✓ **dc:source:**
document_url
- ✓ **dc:type:**
document_type, «mail», «forum», «news», «source», «release», «tracker»
- ✓ **dc:relation:**
ici nous inscrivons si nécessaire un lien vers document_url d'une autre ressource
- ✓ **dc:creator:**
doc_owner_name, mail_sender, forum_author, news_author, source_owner, release_owner, tracker_owner
- ✓ **dc:contributor:**
forge_permission_group_name
- ✓ **dc:publisher:**
forge_project_name

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01

- ✓ **dc:date.created :**
doc_create_date, mail_create_date, forum_date, news_date, source_create_date, srelease_date, tracker_created,
- ✓ **dc:date.modified :**
doc_update_date, source_update_date, tracker_modified,
- ✓ **dc:format:**
file_name_extention
- ✓ **dc:langue:**
«fr», «en»

Tout les métadonnées des Forges qui ne seront pas présentées par Dublin Core seront présentées par un extension de Dublin Core.

Exemple: Nous avons un document Solr qui contient les données suivant dans l'index pour le forge Codendi.

```

<doc>
  <field name="id:268">/field>
  <field name="comments"></field>
  <field name="doc_create_date">[B@84de3c</field>
  <field name="doc_create_day">26</field>
  <field name="doc_create_month">11</field>
  <field name="doc_create_year">2009</field>
  <field name="doc_description">htperf automation (autobench), vizualization, other tools. Taken from
    http://railslab.newrelic.com/2009/06/23/episode-16-load-testing-part-2</field>
  <field name="doc_owner">218</field>
  <field name="doc_owner_name">nterray</field>
  <field name="doc_title">Load testing screencast - part 2</field>
  <field name="doc_update_date">[B@11a47df</field>
  <field name="doc_update_day">26</field>
  <field name="doc_update_month">11</field>
  <field name="doc_update_year">2009</field>
  <field name="document_url">/plugins/docman/?group_id=110&action=details&id=268</field>
  <field name="file_address">/var/lib/codendi/docman/codendiorg/6/8/268/0/file</field>
  <field name="forge_permission_group">2</field>
  <field name="forge_permission_group">3</field>
  <field name="forge_permission_group">4</field>
  <field name="forge_permission_group_name">2</field>
  <field name="forge_permission_group_name">3</field>
  <field name="forge_permission_group_name">4</field>
  <field name="forge_project_id">110</field>
  <field name="forge_project_name">Codendi.org</field>
  <field name="recordtype">Coclico</field>
</doc>


```

D'après notre description de correspondance nous allons obtenir la description des metadonnées en dublin core suivante:

```

<dc>Title>Load testing screencast - part 2</dc>Title>
<dc:Creator>nterray</dc:Creator>
<dc:Subject>Metadata, Dublin Core element, resource description</dc:Subject>
<dc:Description>htperf automation (autobench), vizualization, other tools. Taken from
  http://railslab.newrelic.com/2009/06/23/episode-16-load-testing-part-2
</dc:Description>
<dc:Publisher>Codendi.org</dc:Publisher>
<dc:Identifier>/plugins/docman/?group_id=110&action=details&id=268</dc:Identifier>
<dc:Contributor>2</dc:Contributor>
<dc:Contributor>3</dc:Contributor>
<dc:Contributor>4</dc:Contributor>

```

	Titre du document : Étude format d'échange compatible OAI-PMH
	Référence : L2.6.1
	Version du 0.01

<dc:Date.Created>2009-11-26</dc:Date>

<dc:Date.Modified>2009-11-26</dc:

2.3 Quelques outils open-source pour construire portail de fédération des Forge compatible OAI-PMH?

Solr sera utilisé pour stocker les métadonnées de forges dans son index. Solr est un serveur de recherche pour entreprises, open source et basé sur la librairie Java de recherche Lucene, avec des APIs XML/HTTP, le principe de cache, de réplication, et une interface d'administration Web.

Rapid Visual OAI Tool (RVOT) est utilisé pour construire graphiquement un entrepôt OAI-PMH à partir d'une collection de fichiers. Les dossiers de la collection d'origine peuvent être dans l'un des format acceptable. Le format pris en charge actuellement sont RFC1807, sous-ensemble des formats Marc & COSATI, qui sont des formats de méta-données pour les bibliographies, et le rapport technique. RVOT contribue à définir la correspondance visuelle à partir d'un format natif au format oai_dc, et une fois cela fait l'outil peut répondre aux requêtes OAI-PMH. L'outil est autonome, il est livré avec un serveur web léger et gestionnaire de requête OAI-PMH et est écrit en Java. La conception de RVOT est telle qu'elle peut être facilement étendu pour supporter les autres formats de métadonnées et nous allons les étendre vers un index de Solr.