



Guide d'interopérabilité OAI-PMH pour un référencement des documents numériques dans Gallica

Ce document est destiné aux bibliothèques souhaitant faire apparaître le signalement de leurs documents numérisés dans Gallica, en utilisant le protocole OAI-PMH.

Contacts BnF :
Guillaume Godet (département de la
Coopération)
guillaume.godet@bnf.fr
Christophe Renard (département des
Systèmes d'information)
christophe.renard@bnf.fr
Dominique Stutzmann (département de
l'Information bibliographique et numérique)
dominique.stutzmann@bnf.fr

1. Qu'est-ce que le protocole OAI-PMH ?

OAI-PMH est le sigle de *Open archives initiative - protocol for metadata harvesting*, ce qui signifie "protocole pour la collecte de métadonnées de l'Initiative pour les Archives ouvertes"

Le protocole OAI-PMH est un moyen **d'échanger sur Internet des métadonnées entre plusieurs institutions, afin de multiplier les accès aux documents numériques**. Il permet d'accroître la visibilité des collections numériques sur Internet, de reconstituer virtuellement des corpus à partir de ressources accessibles sur différents sites, d'alimenter des portails thématiques.

Son utilisation est libre, tout comme ses spécifications, disponibles sur le site www.openarchives.org

L'OAI-PMH définit deux types d'acteurs : les **fournisseurs de données**, qui déposent leurs métadonnées sur un serveur web appelé "entrepôt", et les **fournisseurs de service** qui collectent (on dit aussi "moissonnent") ces données pour les intégrer à l'index de leurs propres bibliothèques numériques. Un même établissement peut jouer les deux rôles, diffuser ses métadonnées et collecter celles des autres.

Le fonctionnement de base du protocole OAI-PMH repose sur une **communication de client à serveur**. Le client envoie des requêtes au serveur en http, le serveur répond par un flux de données en XML.

2. Comment mettre en place un entrepôt OAI-PMH ?

Pour utiliser OAI-PMH dans le cadre des bibliothèques numériques, il faut disposer d'une part de **documents numérisés accessibles en ligne**, d'autre part de **descriptions** de ces documents.

Le principe est de rendre ces descriptions accessibles et récupérables ("moissonnables") sur un serveur web appelé aussi "entrepôt". Chaque description de document doit inclure **une URL pointant vers le document décrit**, dans sa version numérique.

Tout document doit être décrit selon le format de métadonnées généralistes **Dublin Core non qualifié**. On peut également fournir une description du même document dans un autre format, pourvu qu'il soit encodé en XML.

Le format Dublin Core non qualifié définit 15 éléments de métadonnées, dont chacun est facultatif et répétable. Voir : <http://dublincore.org/> et ci-dessous.

Les notices au format Dublin Core non qualifié peuvent être créées manuellement, ou bien obtenues après conversion à partir d'une base de données existantes (catalogues, GED etc.). Ce dernier cas est le plus fréquent.

A partir des informations descriptives dont on dispose, on étudie comment elles seront exprimées dans le format Dublin Core non qualifié.

3. En terme de moyens informatiques, que représente la mise en place d'un entrepôt OAI-PMH ?

Du point de vue informatique, la création d'un entrepôt OAI-PMH nécessite les composants suivants :

- **des métadonnées** sur les ressources stockées dans une base de données (par exemple une base SQL),
- **un serveur web**, accessible par Internet (Apache, IIS...),
- **une interface de programmation / API** (Perl, PHP, Java-Servlet...),
- **un identifiant pour l'entrepôt**, l'URL de base,
- **une application capable de répondre aux 6 requêtes OAI** (Identify, ListSets, ListIdentifiers, ListMetadataFormats, ListRecords, GetRecord)
- pour les entrepôts OAI contenant plusieurs milliers de notices, **une gestion du flux** permettant de renvoyer les notices par paquets.

Dans la plupart des cas, les métadonnées disponibles sur un entrepôt OAI sont obtenues par conversion depuis une base de données existante. L'entrepôt OAI ne remplace pas la base d'origine, mais en constitue une extension.

De nombreuses **solutions open source** sont disponibles en ligne, dont la liste figure sur le site officiel de l'Open Archives Initiative
<http://www.openarchives.org/pmh/tools/tools.php>

La BnF utilise pour ces deux entrepôts OAI le logiciel open source développé par l'OCLC, appelé, OAIcat, disponible à l'URL suivant :
<http://www.oclc.org/research/software/oai/cat.htm>

Du côté de l'**offre commerciale**, on se référera à l'article "Les progiciels métier disponibles sur le marché en 2008, leur couverture fonctionnelle et leur cadre technique", Livres Hebdo n°123 du vendredi 29 février 2008, page 78 et suivantes.

4. Comment s'assurer que mon entrepôt respecte bien les règles techniques du protocole OAI-PMH ?

Pour tester la validité d'un entrepôt OAI, c'est-à-dire sa conformité aux règles de bases, on peut utiliser un **outil de validation en ligne**. Par exemple,

L'outil de validation Repository Explorer (Virginia Tech / University of Cape Town)
<http://re.cs.uct.ac.za/>

Après s'être assuré de la validité de son entrepôt OAI, on peut le faire connaître en l'inscrivant sur les principaux registres (répertoires en ligne) :

- *Registered data providers* sur le site officiel de l'OAI
<http://www.openarchives.org/Register/BrowseSites>
- *The University of Illinois OAI-PMH Data Provider Registry*
<http://gita.granger.uiuc.edu/registry/>

5. Comment structurer un entrepôt OAI-PMH en sous-ensembles ?

Un **entrepôt OAI** peut être structuré en **sous-ensembles** (en anglais, "sets"). Cette structuration est facultative. Elle permet le moissonnage sélectif sur un ensemble donné de notices, sans avoir à récupérer toutes les notices contenues dans l'entrepôt.

Il est possible de créer des ensembles par **thèmes**, en s'appuyant par exemple sur les grandes classes Dewey, ou bien par **types de documents** (livres, photographies, documents d'archives, périodiques etc.), ou encore par **fonds ou collections**.

Exemple

Il est possible de récupérer uniquement les notices de Gallica

- des ouvrages et périodiques portant sur le droit (set *gallica:3:34*)
 - des photographies et lots de photographies (set *gallica:images:photographies*)
 - des partitions du fonds Philidor (set *gallica:philidor:partitions*)
- etc.

Les sets peuvent être créés dans le cadre de **projets coopératifs** entre plusieurs établissements.

Pour l'intégration dans Gallica et la recherche par types de documents (manuscrits / livres / presse et périodique / partitions / enregistrements sonores / cartes), il est souhaitable que certains sets puissent être identifiés comme correspondant à des types de documents.

Les sets peuvent être hiérarchisés. Dans ce cas, on utilise les deux-points (":") pour signaler la relation d'appartenance.

Exemple :

gallica:images:photographies indique que le set "photographies" appartient à l'ensemble "images", lui-même inclus dans l'ensemble "gallica".

Une même notice peut appartenir à plusieurs sets.

NB : Attention, l'élément `<setSpec>` sert à indiquer le code du set dans la requête HTTP qui permet d'en récupérer les notices. Il ne doit pas comporter d'espace, ni d'accents, ni de caractères spéciaux autres que les neuf caractères suivants :

- _ . ! ~ * ' ()

En revanche, l'élément `<setName>` est descriptif et peut contenir tous types de caractères.

Exemple :

`<setSpec>gallica:anville</setSpec>`

`<setName>Gallica : fonds d'Anville</setName>`

6. Quelles sont les conditions requises, du point de vue documentaire, pour une intégration des notices OAI dans Gallica ?

Pour pouvoir être signalés dans Gallica, les documents numérisés doivent être accessibles en accès libre sur un site web externe, c'est-à-dire qu'ils doivent appartenir au domaine public ou bien faire l'objet d'un accord avec les ayants-droit en vue de leur libre utilisation par les internautes.

Le document ne doit pas avoir été déjà numérisé par la BnF, sauf si l'édition est différente.

Le contenu du document doit correspondre aux principes de la politique documentaire de Gallica.

7. Check-list

Tous les points ci-dessous sont des préalables indispensables pour une intégration dans Gallica.

(N.B. : Toutes les erreurs décrites ont déjà été constatées.)

1. Validité technique

Tester son entrepôt avec l'outil <http://re.cs.uct.ac.za/>

2. Identify

Tester l'entrepôt avec la requête ?verb=Identify

- a) repositoryName : Vérifier que le nom de l'entrepôt est correspond bien à celui de son institution (certains prestataires oublient de mettre à jour les paramètres)
- b) baseURL : Vérifier que l'URL de base est bien la même que celle utilisée pour formuler les requêtes
- c) protocolVersion : Utiliser la version 2.0
- d) adminEmail : Vérifier l'adresse électronique de l'administrateur ou d'un responsable technique. (Une erreur fréquente est de laisser un paramètre du logiciel utilisé)
- e) earliestDatestamp :
 - a. Ne pas confondre « datestamp » et date des ressources
 - b. Vérifier que la date est crédible (si vous créez votre entrepôt en 2010, le plus ancien datestamp est de 2010 !)
 - c. Vérifier que sa forme correspond à celle de la granularité énoncée juste en-dessous, soit AAAA-MM-JJTHH:MM:SSZ (2010-04-27T16:21:54Z), soit AAA-MM-JJ (2010-04-27).
- f) granularity : il s'agit de la précision des dates des enregistrements
Attention : les dates des éléments <responseDate> et <resumptionToken> est toujours de la forme AAAA-MM-JJTHH:MM:SSZ

Si l'entrepôt est décrit avec le schéma <http://www.openarchives.org/OAI/2.0/oai-identifieur.xsd>, il ne faut pas oublier que :

- a) l'élément <scheme> contient oai
- b) l'élément <repositoryIdentifier> *ne contient pas* le préfixe « oai: »
- c) l'élément <delimiter> est obligatoirement le double-point
- d) l'élément <sampleIdentifier> est obligatoire

A SAVOIR : les identifiants OAI

En premier lieu, il est nécessaire d'avoir un **identifiant OAI qui soit spécifique à l'établissement**, et non pas générique.

En particulier, les bibliothèques qui utilisent la solution logicielle d'OCLC veilleront à ne pas laisser l'identifiant par défaut qui est :

```
<repositoryIdentifier>oaicat.oclc.org</repositoryIdentifier>  
<sampleIdentifier>oai:oaicat.oclc.org:OCLCNo/ocm00000012</sampleIdentifier>
```

Elles doivent le remplacer par leur propre identifiant.

Cet identifiant servira à indiquer dans Gallica, pour chaque notice moissonnée, son établissement de provenance.

L'identifiant des enregistrements dans un entrepôt OAI est de la forme :

oai:nomdedomaine.fr:identifiant

constitué de :

- préfixe « oai »
- double-point « : »
- nom de domaine, commençant par une lettre, avec une séquence de chiffres et lettres (sans espace, ni caractère accentué ou spécial, sauf le tiret) et comportant au moins un point « . » avant une nouvelle séquence de lettres et chiffres (sans espace, ni caractère accentué ou spécial, sauf le tiret)
- double-point « : »
- identifiant de l'enregistrement, composé de chiffres et de lettres et pouvant comprendre les caractères spéciaux suivants :
- _ . ! ~ * ' () ; / ? : @ = + \$, %

Cf. <http://www.openarchives.org/OAI/2.0/guidelines-oai-identifieur.htm>

3. ListSets

Vérifier que :

- les éléments `<setSpec>` ne doivent pas comporter d'espace, ni d'accents, ni de caractères spéciaux autres que les neuf caractères suivants :
- _ . ! ~ * ' ()

4. ListMetadataFormats

La réponse doit au moins comprendre la section suivante :

```
<metadataFormat>  
  <metadataPrefix>oai_dc</metadataPrefix>  
  <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema>  
  <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc</metadataNamespace>  
</metadataFormat>
```

5. ListIdentifiers + from-until + set

Vérifier les points suivants :

- `<identifiant>`
 - o Si le schéma <http://www.openarchives.org/OAI/2.0/oai-identifiant.xsd> est utilisé dans la requête Identify, les identifiants sont de la forme :
 - oai:domaine.xx:identifiantlocal
 - o Il est toujours conseillé d'éviter les caractères spéciaux
- `<datestamp>`
 - o La granularité correspond bien à celle annoncée dans la réponse à la requête Identify
 - o La date de l'enregistrement le plus ancien est exacte
 - La date annoncée dans `<earliestDatestamp>` est souvent plus reculée que la date réelle. Il faut donc parfois tâtonner avec les paramètres 'from' et 'until'
- `<setSpec>`
 - o Le set spécifié dans la requête doit apparaître dans l'en-tête des enregistrements

6. ListRecords et section <metadata>

Concernant les éléments du Dublin Core simple requis pour Gallica, on distingue deux niveaux :

- les **éléments obligatoires**, dont l'absence ou le mauvais usage constitueraient des points bloquants pour le signalement des documents dans Gallica
- les **éléments facultatifs**, dont l'usage permettra d'exploiter au mieux les possibilités d'indexation du moteur Gallica

Le respect des normes internationales et nationales s'applique aussi bien pour la construction des vedettes que pour la transcription des zones de titre

| Élément DC | statut | Conditions pour une prise en charge dans Gallica |
|----------------|------------------------|---|
| dc:title | obligatoire | <p>Ponctuation ISBD sauf "indication générale du type de document".</p> <p>Nota bene :</p> <ul style="list-style-type: none"> - Le sous-titre fait partie de l'élément <dc:title> avec ponctuation ISBD (point entre titre propre et titre dépendant) - l'ISSN ne fait pas partie du titre d'un périodique - pour la description des fascicules de périodiques à l'unité, le numéro peut être ajouté dans le titre |
| dc:creator | obligatoire (si connu) | <p>Norme Afnor NF Z 44-06 1 => Nom, Prénom (date-date)</p> <p>Une personne = un élément</p> <ul style="list-style-type: none"> - Une seule personne (morale ou physique) par élément - Autant d'éléments répétés que d'auteurs reconnus <p>Normaliser les entrées</p> <ul style="list-style-type: none"> - dans un même entrepôt, on ne doit pas trouver « Dupont, M. », « Dupont, Martin », « Martin Dupont » s'il s'agit d'une seule et même personne ! <p>Ne relever que les auteurs de la ressource décrite <i>NE PAS utiliser cet élément pour :</i></p> <ul style="list-style-type: none"> - les provenances et anciens collectionneurs ne sont pas des <dc:creator> - les auteurs d'un fonds - les architectes d'un bâtiment photographié (l'auteur est le photographe, pas l'architecte !) - les sujets (personnes photographiées ou sujets de biographies) |
| dc:contributor | facultatif | Mêmes conditions que pour <dc:creator> |
| dc:publisher | obligatoire (si connu) | Sous la forme suivante : éditeur (ville) |

| Élément DC | statut | Conditions pour une prise en charge dans Gallica |
|----------------|------------------------|--|
| dc:date | obligatoire (si connu) | Utiliser ISO 8601 [W3CDTF], c'est-à-dire la forme AAAA-MM-JJ <ul style="list-style-type: none"> - Année en quatre chiffres - Période avec la ponctuation AAAA/AAAA ou AAAA-MM-JJ/AAAA-MM-JJ - Utiliser « 14.. » et non « 15e siècle » - Ne pas utiliser de crochets carrés, ni de préfixe « ca. ». Les incertitudes sur la datation sont à mettre en note en <dc:description> |
| dc:language | facultatif | S'il est utilisé, respecter la norme ISO 639-2b (code pour la représentation des noms de langues) sur 3 caractères Ex. <dc:language>fre</dc:language> pour un document en langue française |
| dc:description | facultatif | |
| dc:coverage | facultatif | Le périmètre temporel s'exprime par des dates avec les mêmes conditions que <dc:date> <ul style="list-style-type: none"> - éviter « 19s » <p>Le périmètre spatial s'exprime en localisations non ambiguës et autonomes, selon la construction Rameau</p> <ul style="list-style-type: none"> - Au lieu de : <dc:coverage>Hoche, rue</dc:coverage> écrire : <dc:coverage> Roubaix (Nord) – Rue Hoche</dc:coverage> |

| Élément DC | statut | Conditions pour une prise en charge dans Gallica |
|------------|-------------|--|
| dc:subject | facultatif | <p>Pour les fonds locaux, rendre l'élément <dc:subject> indépendant du reste de la notice</p> <ul style="list-style-type: none"> - Au lieu de : <dc:coverage>Hoche, rue</dc:coverage> écrire : <dc:coverage> Roubaix (Nord) – Rue Hoche</dc:coverage> <p>Ne relever que les sujets de la ressource décrite <i>NE PAS utiliser cet élément pour :</i></p> <ul style="list-style-type: none"> - les provenances et anciens collectionneurs ne sont pas des <dc:subject> - les auteurs d'un fonds - les architectes d'un bâtiment photographié (l'auteur est le photographe, le sujet est le bâtiment, pas l'architecte !) - |
| dc:format | obligatoire | <p>Respecter les types MIME Ex. <dc:format>image/jpeg</dc:format> ou <dc:format>application/pdf</dc:format></p> <p>Un type MIME = Un élément Si plusieurs formats sont disponibles ou s'il faut donner plusieurs informations, répéter l'élément</p> <ul style="list-style-type: none"> - préférer : <dc:format>image/jpeg</dc:format> <dc:format>388 Mo </dc:format> - à <dc:format>image/jpeg ; 388 Mo</dc:format> <p>Format de la source C'est le <dc:format> qui permet de décrire le document physique originel (nombre de pages, dimensions etc.)</p> |

| Élément DC | statut | Conditions pour une prise en charge dans Gallica |
|---------------|-------------|---|
| dc:type | obligatoire | <p>Créer au moins un <dc:type> comprenant l'une des valeurs du référentiel DCMI http://dublincore.org/documents/dcmi-type-vocabulary/</p> <p>Le <dc:type> doit être répété avec des valeurs plus précises, comme celles actuellement prises en charge dans Gallica :</p> <ul style="list-style-type: none"> - « monographie imprimée » - « publication en série imprimée » - « image fixe » - « document cartographique » - « musique imprimée » - « enregistrement sonore » |
| dc:identifier | obligatoire | <p>Un élément <dc:identifier> doit comporter le lien externe vers le document</p> <p>Un identifiant = Un élément Les ISBN, ISSN et autres identifiants sont à inscrire dans cet élément répétable</p> |
| dc:relation | facultatif | A utiliser pour décrire les rapports de tout à partie |
| dc:source | facultatif | <p>comporte les informations sur le document d'origine, en particulier :</p> <ul style="list-style-type: none"> - la cote de l'exemplaire physique de façon autonome et non ambiguë <p><u>Exemples</u></p> <ul style="list-style-type: none"> - Bibliothèque nationale de France, 8-Ye-1701 (et non : « 8-Ye-1701 ») - Roubaix, Médiathèque, MS 008 (et non « MS_008 ») <p><u>Nota bene</u> : en Dublin Core simple, il est <i>interdit</i> d'utiliser l'élément <provenance> !</p> |
| dc:rights | obligatoire | Ex. « domaine public »/ « public domain » |

8. Préconisations supplémentaires pour une intégration des notices OAI dans Gallica

Le protocole OAI-PMH prévoyant un moissonnage et une présentation asynchrone des résultats, il faut s'assurer de **la stabilité des URL** pointant vers les documents numériques. Cette stabilité peut être assurée par le recours à un résolveur de liens. Cette solution permet de communiquer une URL pérenne, indépendamment de la localisation du document sur le serveur qui peut changer dans le temps.

L'**infrastructure réseau** doit être correctement dimensionnée et il est nécessaire d'avoir un serveur qui accepte **plusieurs accès simultanés**, et qui tienne la charge pendant le moissonnage.

Enfin, il sera utile d'indiquer la **fréquence de mise à jour et d'alimentation** de l'entrepôt OAI (quotidienne/hebdomadaire/mensuelle/trimestrielle etc.), afin d'optimiser le moissonnage des données par Gallica.

9. Où puis-je trouver de l'information sur les entrepôts OAI-PMH de la BnF ?

Les informations concernant les deux entrepôts OAI de la BnF sont disponibles sur le site **bibnum.bnf.fr** à la rubrique OAI. On y trouvera notamment les URL de base des entrepôts, des informations concernant les mises à jour et l'enrichissement de ces réservoirs de données, les formats de métadonnées utilisés et le *Guide d'utilisation du Dublin Core non qualifié à la BnF*.

9. Comment moissonner les métadonnées provenant de Gallica ou d'autres bibliothèques numériques ?

La récupération de notices OAI s'effectue par le biais d'un logiciel dit "**moissonneur**" (en anglais, *harvester*). Les notices sont ensuite stockées localement dans une base de données.

Il existe plusieurs logiciels moissonneurs en open source, dont la liste est disponible à l'URL :

<http://www.openarchives.org/pmh/tools/tools.php>