

Moissonnage des données

Exposé général sur les principes

Nader Boutros,

PASS-TECH, EVCAU, FMSH

nader (at) boutros (dot) org



membre de l'association **ADNX**

Association pour la Documentation Numérique en XML



Plan

- ⇒ Archives électroniques ouvertes
- ⇒ Le mouvement "open access"
- ⇒ OAI-PMH
 - Protocole pour le moissonnage des données
- ⇒ Fournisseur de données
- ⇒ Fournisseurs de services



Archives électroniques ouvertes

- ⇒ Pourquoi ?
 - Pour garantir l'accès à la littérature de recherche
 - Pour accélérer le rythme de publication
- ⇒ Pour qui ?
 - Pour les chercheurs qui ont besoin de s'informer des travaux de leurs collègues le plus tôt possible
- ⇒ Qui l'a adopté ?
 - Les chercheurs en médecine, physique, mathématiques, sciences de l'information et de la communication ... puis les autres suivent.

 Voir dossier sur l'histoire des archives ouvertes par Hélène BOSCH, bibliothécaire INRA Tours
<http://cogprints.org/4408/01/OuvrageOAIarchive.pdf> + [page web qu'elle maintient](#)



Le mouvement Open Access

ou libre accès

- ⇒ Démocratisation de l'accès à la littérature de recherche (articles, thèses, ...) depuis 1991 :
 - gratuité + texte intégral
 - Déclaration de Budapest 2002, Berlin1 2003, Berlin2 2004 et Berlin3 à Southampton en 2005
- ⇒ Serveurs de services d'archivage en ligne
 - Garantir le libre accès
 - Diffusion sans restriction (OAI)



Voir dossier sur l'Open Access préparé par :
SERVICE DOCUMENTAIRE ECOLE NATIONALE DES PONTS ET CHAUSSEES
http://www.enpc.fr/fr/documentation/doc_electronique/open_access_enpc05.pdf



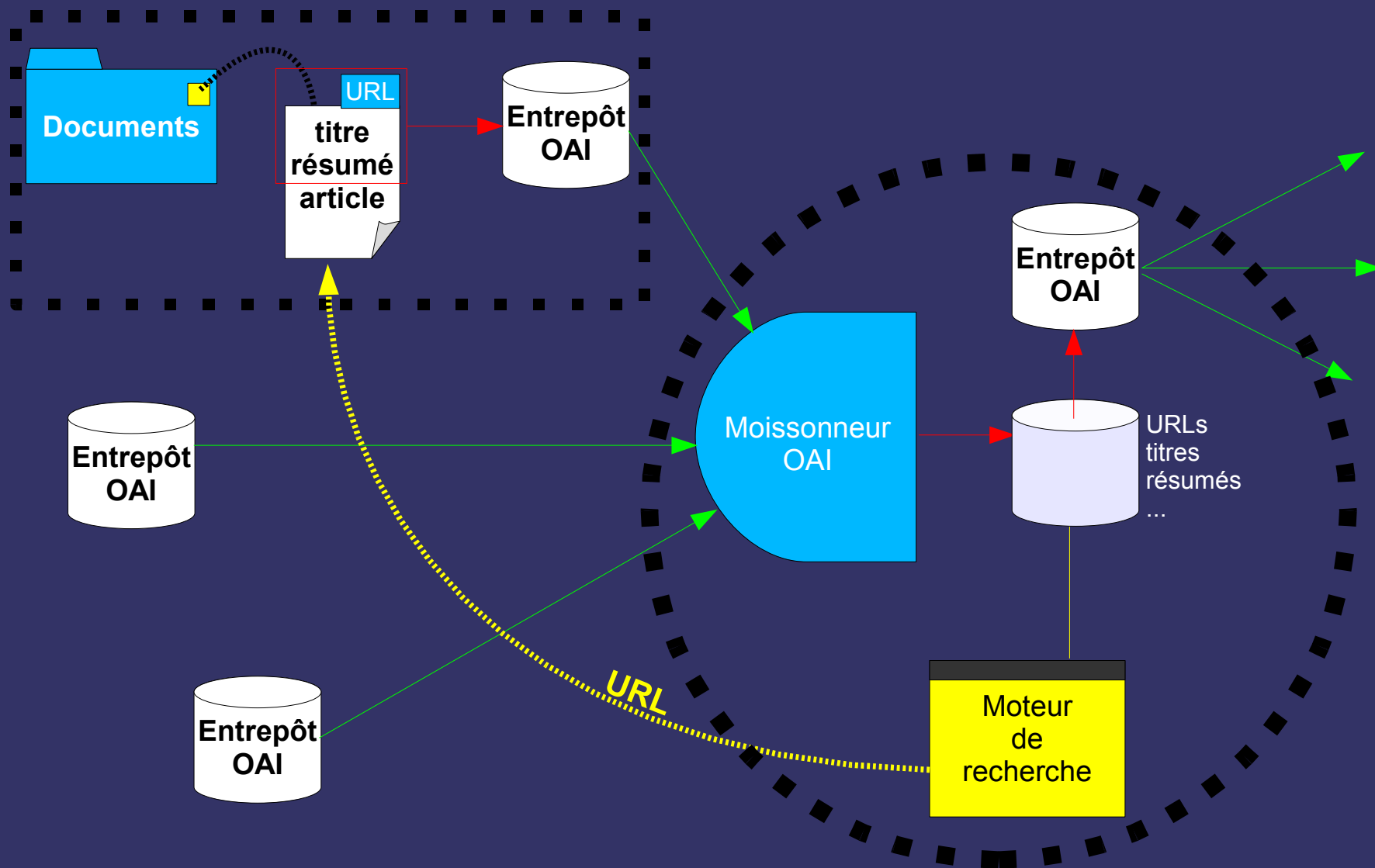
OAI-PMH : Modèle

Open Archives Initiative Protocol for Metadata Harvesting

- ⇒ Accès à des sources hétérogènes de données d'une manière homogène!
- ⇒ Deux types d'acteurs :
 - Les entrepôts exposent des métadonnées en implémentant le protocole OAI
 - Les moissonneurs recueillent des métadonnées à l'aide du protocole OAI
- ⇒ Moissonner =
 - récupérer une copie des métadonnées en local puis la rendre cherchables comme valeur ajoutée
 - Un entrepôt sert plusieurs moissonneurs
 - Un moissonneur moissonne plusieurs entrepôts
 - Exemple : **OAIster** Voir : <http://www.oaforum.org>

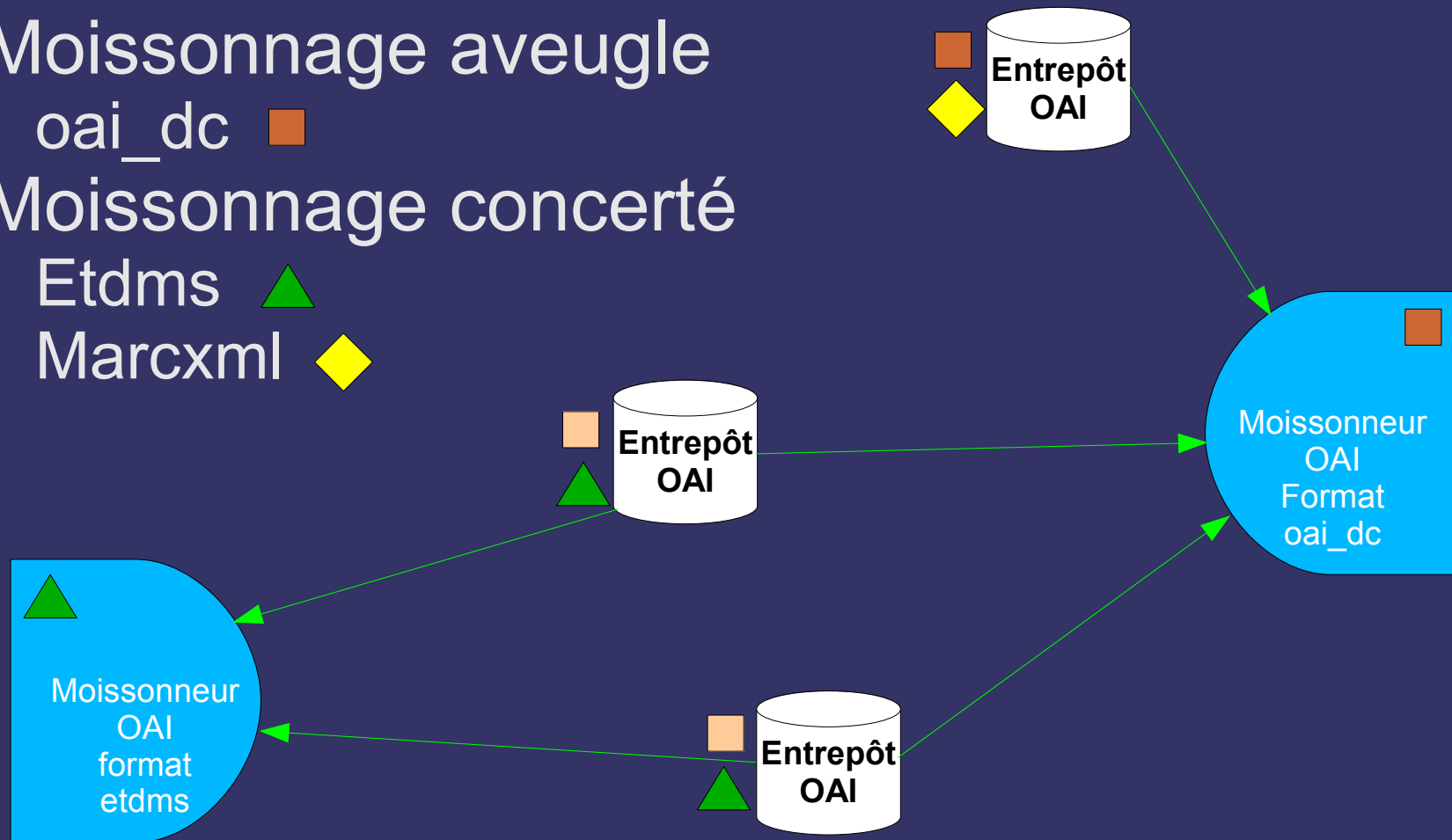


OAI-PMH : Organisation



Types de moissonnage

- ⇒ Moissonnage aveugle
 - oai_dc ■
- ⇒ Moissonnage concerté
 - Etdms ▲
 - Marcxml ◆



OAI-PMH : Spécification

Open Archives Initiative Protocol for Metadata Harvesting

- ⇒ Protocole pour le moissonnage des données :
 - garantir l'interopérabilité
- ⇒ Indépendance des logiciels et systèmes
- ⇒ Resource (thèse, article, livre de bibliothèque)
 - Item (unique ID)
description de la ressource par un ensemble de données descriptives
 - Record (format 1)
 - Record (format 2)
- ⇒ Flux XML par HTTP
- ⇒ Requête URL = réponse XML
- ⇒ Format obligatoire + formats personnalisés

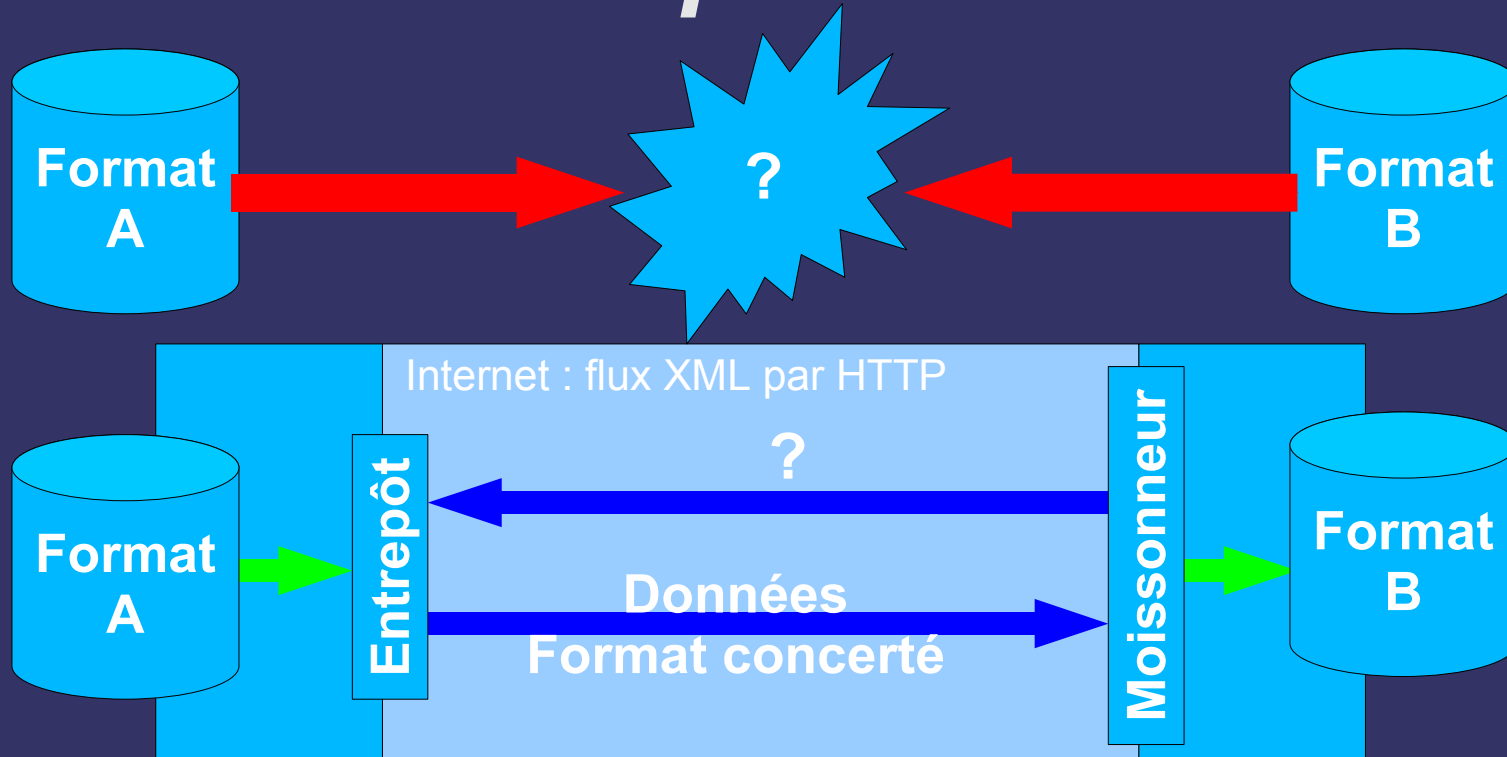


Voir les spécifications du protocole :

<http://www.openarchives.org/OAI/openarchivesprotocol.html>



OAI-PMH : Entrepôt et moissonneur



- ⇒ Un entrepôt est par définition passif
 - Ne répond qu'aux demandes
- ⇒ Un moissonneur est par définition actif
 - Interroge les entrepôts à moissonner

OAI-PMH : URL HTTP

- ⇒ Construire une URL (adresse web)
- ⇒ Adresse d'un entrepôt :
 - <http://mon-entrepot-oai.org/outil-oai>
- ⇒ Premier paramètre obligatoire est :
 - verb=[1 des 6 verbes]
- ⇒ Autres paramètres obligatoires ou optionnels
- ⇒ Exemple de la BnF :
 - **Enregistrements** oai_dc du 1/11 au 1/12/2005
[http://oai.bnf.fr/repositoryOAI.php
?verb=ListRecords
&metadataPrefix=oai_dc
&from=2005-11-01
&until=2005-12-01](http://oai.bnf.fr/repositoryOAI.php?verb=ListRecords&metadataPrefix=oai_dc&from=2005-11-01&until=2005-12-01)



OAI-PMH : Verbes

- ⇒ Identify : l'identité de l'entrepôt OAI
- ⇒ ListMetadataFormats : la liste des formats
- ⇒ ListSets : la liste des collections
- ⇒ ListIdentifiers : la liste des identifiants
- ⇒ ListRecords : les enregistrements
- ⇒ GetRecord : un enregistrement



OAI-PMH : XML / HTTP

➔ Exemple : <http://arxiv.org/oai2?verb=ListMetadataFormats>

```
➔ <OAI-PMH xsi:schemaLocation="
    http://www.openarchives.org/OAI/2.0/
    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2005-11-27T08:40:16Z</responseDate>
  <request verb="ListMetadataFormats">
    http://arXiv.org/oai2
  </request>
  <ListMetadataFormats>
    <metadataFormat>
      <metadataPrefix>oai_dc</metadataPrefix>
      <schema>http://www.openarchives.org/OAI/2.0/oai_dc
      <metadataNamespace>http://www.openarchives.org/OAI
    </metadataFormat>
  </ListMetadataFormats>
</OAI-PMH>
```



OAI-PMH : Formats

- ⇒ Schéma fourni par OAI pour :
 - Dublin Core non qualifié (oai_dc) !! **Obligatoire**
 - <http://dublincore.org/documents/dces/>
- ⇒ Autres formats :
 - MARC <http://www.loc.gov/marc/>
 - MACHine-Readable Cataloging
 - MODS <http://www.loc.gov/standards/mods/>
 - Metadata Object Description Schema
 - ETD-MS <http://www.ndltd.org/standards/metadata/current.html>
 - Electronic Theses and Dissertations Metadata Standard
 - OLAC <http://www.language-archives.org/OLAC/metadata.html>
 - Open Language Archives Community
 - rfc1807, marcxml, eprints, ...
 - Format de donnée propre à un projet



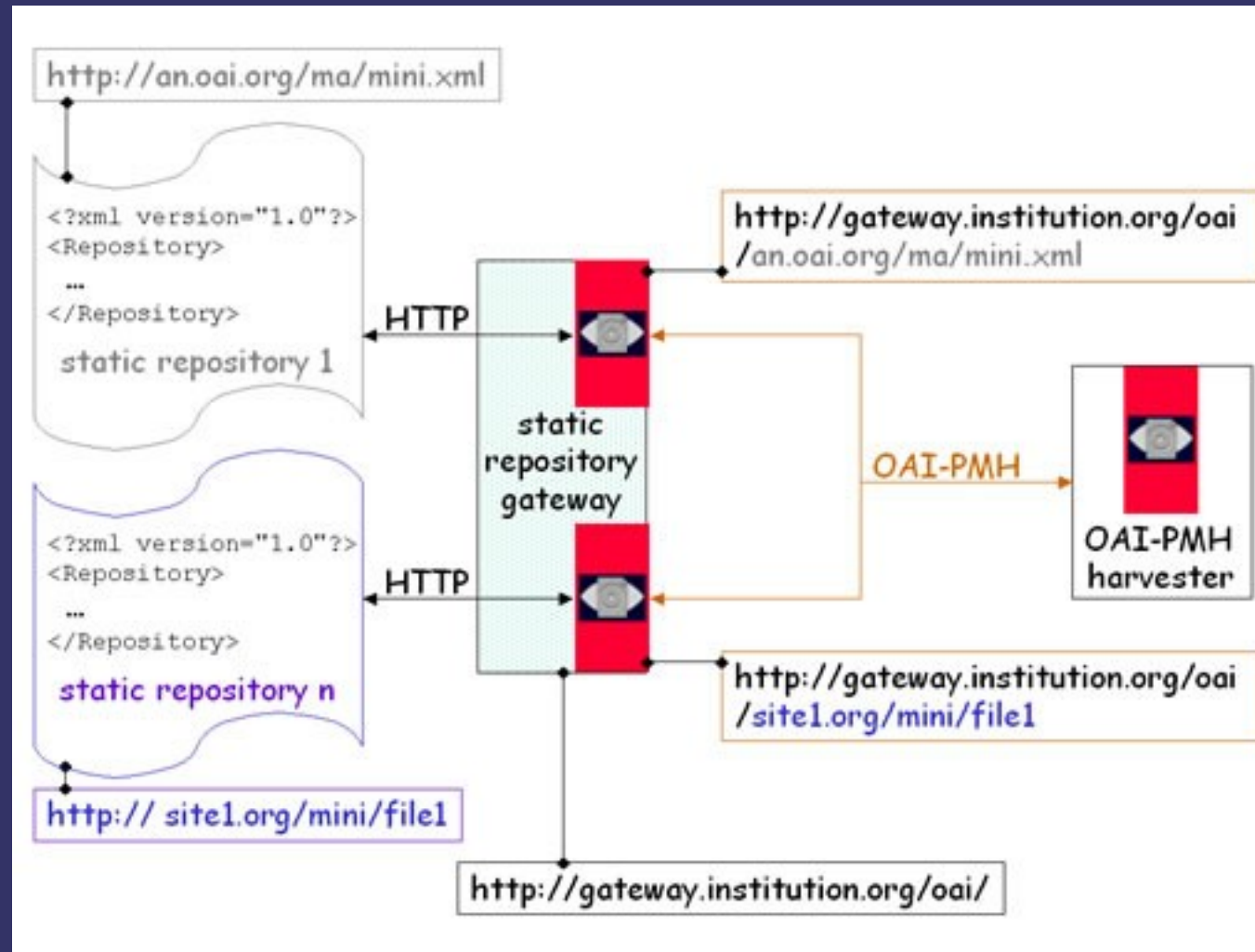
OAI-PMH : Dublin Core non-qualifié

dc:identifiant	Identifiant
dc:title	Titre
dc:subject	Sujet
dc:description	Description du contenu
dc:date	Une date significative du contenu
dc:coverage	Couverture spatiale ou temporelle
dc:creator	Responsable du contenu
dc:contributor	Responsable secondaire du contenu
dc:publisher	Responsable de la diffusion
dc:type	Nature du contenu (text, image, ...)
dc:format	Format du contenu (pdf, jpeg, ...)
dc:language	Langue du contenu
dc:source	Identifiant d'une autre ressource sur le contenu
dc:relation	Référence à un autre contenu
dc:rights	Copyright



Types d'entrepôts

- ➔ Entrepôts OAI standard
- ➔ Entrepôts OAI statique



Extrait des spécifications d'entrepôt et passerelle d'entrepôts OAI statiques : <http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm>



Fournisseurs de données

- ⇒ Bibliothèques
- ⇒ Centres de recherche/universités
- ⇒ Journaux
- ⇒ Editeurs
- ⇒ Projets divers
- ⇒ ...
- ⇒ Et bientôt vous!



Fournisseurs de données

⇒ Bibliothèques

- Bibliothèque du congrès américain (LOC) :
http://memory.loc.gov/cgi-bin/oai2_0?verb=Identify
- La BnF : Bibliothèque Nationale de France
<http://oai.bnf.fr/repositoryOAI.php?verb=Identify>

⇒ Centres de recherche/universités

⇒ Journaux

⇒ Editeurs

⇒ Projets divers



Fournisseurs de données

⇒ Bibliothèques

⇒ Centres de recherche/universités

- **Thèses en ligne (univ. Lyon2) : Cyberthèses**

<http://demeter.univ-lyon2.fr:8080/sdx/sdx/oai/theses/documents?verb=Identify>

- **CNRS : CCSd SIC / HAL / TEL :**

<http://archivesic.ccsd.cnrs.fr/perl/oai20?verb=Identify>

<http://hal.ccsd.cnrs.fr/oai/oai.php?verb=Identify>

<http://tel.ccsd.cnrs.fr/oai/oai.php?verb=Identify>

- **CogPrints : <http://cogprints.soton.ac.uk> - Identify**

- **ArXiv : <http://arxiv.org> - Identify**

⇒ Journaux

⇒ Editeurs

⇒ Projets divers



Fournisseurs de données

- ⇒ Bibliothèques
- ⇒ Centres de recherche/universités
- ⇒ **Journaux**
 - **AJOL** : African Journals Online
<http://www.ajol.info/oai?verb=Identify>
- ⇒ Editeurs
- ⇒ Projets divers



Fournisseurs de données

- ⇒ Bibliothèques
- ⇒ Centres de recherche/universités
- ⇒ Journaux
- ⇒ **Editeurs**
 - <http://romeo.eprints.org/publishers.html>
 - Elsevier : <http://www.sciencedirect.com>
 - Ex : <http://dx.doi.org/10.1016/j.prps.2004.07.005>
 - Ulrichsweb : <http://www.ulrichsweb.com>
- ⇒ Projets divers



Fournisseurs de données

- ⇒ Bibliothèques
- ⇒ Centres de recherche/universités
- ⇒ Journaux
- ⇒ Editeurs
- ⇒ Projets divers
 - Strabon - <http://www.strabon.org>
 - Michael - <http://www.michael-culture.org>
 - ...



Fournisseurs de services

- ⇒ Oaister : <http://oaister.umdl.umich.edu/o/oaister/>
- ⇒ Eprints : <http://archives.eprints.org>
- ⇒ DOAJ : <http://www.doaj.org>
- ⇒ Cite-base : <http://www.citebase.org>
- ⇒ ARC : <http://arc.cs.odu.edu>
- ⇒ Citeseer : <http://citeseer.ist.psu.edu>
 - Un graphe ? <http://www.pmbrowser.info/citeseer.html>
- ⇒ Mais on peut aussi développer son propre service de recherche dans les métadonnées de plusieurs archives sélectionnées.
- ⇒ Démo en local : sdx - xtogen et pkp harvester



Outils libres

⇒ Entrepôts / Réservoirs / *Repositories*

- GNU eprints : <http://software.eprints.org/>
- D-space : <http://www.dspace.org/>
- PHPOAI2 : <http://physnet.uni-oldenburg.de/oai>
- SREPOD : <http://srepod.sourceforge.net>
- Framework SDX : <http://adnx.org/sdx>
- XToGen : <http://xtogen.tech.fr>

⇒ Moissonneurs / *Harvesters*

- PKP Harvester : <http://pkp.sfu.ca/pkp-harvester/>
- OAI/ODL Harvester : <http://oai.dlib.vt.edu/odl/software/harvest>
- Framework SDX : <http://adnx.org/sdx>



Conclusion

- ⇒ Open archives
 - ⇒ Open Access
 - ⇒ Open your mind ;-)
-
- ⇒ Cette présentation est en ligne :
 - <http://tech.fr/oai/harvesting.pdf>

