Le protocole OAI-PMH et les fonctionnalités de recherche : étude de portails du domaine patrimonial

Mélanie Gauthier

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

INSTITUT NATIONAL DES TECHNIQUES DE LA DOCUMENTATION

MEMOIRE pour obtenir le

Titre professionnel « Chef de projet en ingénierie documentaire » INTD

niveau I

présenté et soutenu par

Mélanie Gauthier

le 2 novembre 2007

Le protocole OAI-PMH et les fonctionnalités de recherche

Etude de portails du domaine patrimonial

Jury
Séverine Denys
Patrice Verrier

Cycle supérieur Promotion XXXVII

A ma famille, à mes amis

Remerciements

Mes remerciements vont tout d'abord à Patrice Verrier et à ses collaborateurs du « G8 » à la Cité de la Musique pour leur accueil chaleureux et leur disponibilité tout au long de mon stage.

Merci à Séverine Denys pour ses conseils.

Merci à Rodolphe Bailly, à Frédéric Martin et à Isabelle Gauchet pour le temps qu'ils m'ont consacré.

Merci à mes amies Julia et Elise pour leur soutien.

Et enfin, merci à mon amie Soraya, pour ses nombreuses illustrations du mot solidarité !

Notice

GAUTHIER Mélanie. Le protocole OAI-PMH et les fonctionnalités de recherche : étude de portails du domaine patrimonial. 2007. 135 p. Mémoire de diplôme supérieur, INTD-CNAM.

Le protocole OAI-PMH a été adopté rapidement après son apparition par de nombreuses institutions culturelles désireuses de créer des ponts entre leurs collections et de se rendre plus visibles sur le web par l'intermédiaire de portails permettant l'accès à leurs ressources disséminées. Cependant, cette mise en commun dans des index centralisés de métadonnées de provenance et de type divers pose des questions en termes de recherche. Le mémoire a pour objectif d'identifier les problèmes les plus fréquemment rencontrés et de fournir des préconisations pour y remédier.

Bonne pratique ; Evaluation ; Interopérabilité ; Métadonnée ; OAI-PMH ; Patrimoine culturel numérique ; Portail internet ; Recherche en ligne ; Site web

Table des matières

Liste des tableaux

Liste des figures

Avant-propos

Voici trois **codes** utilisés dans le mémoire dans le but d'en faciliter la lecture :

- Les **numéros** suivis de **noms d'auteurs** présentés entre parenthèses (2, Arms) renvoient aux références de la **bibliographie** thématique analytique (p. 90) ;
- Les mots suivis d'un **astérisque** sont définis dans le **glossaire** (p. 101) ;
- Les termes *fournisseur de données, fournisseur de services* et *Dublin Core* revenant souvent, les initiales **FD**, **FS** et **DC** sont utilisées.

La littérature sur le sujet étant très majoritairement publiée en anglais, certains termes essentiels sont **traduits** en **anglais**. Ils apparaissent alors en italique.

Introduction

Il se développe depuis le début des années 2000 de nombreux projets fondés sur l'utilisation d'un protocole récemment apparu : l'**OAI-PMH**. L'Open Archives Initiative Protocol for Metadata Harvesting (**protocole*** de collecte de **métadonnées*** de l'Initiative des Archives Ouvertes) propose un **standard** relativement simple pour l'échange de métadonnées. Créé par l'Initiative des Archives Ouvertes en 1999, il avait pour objectif premier de « *faciliter la description et la diffusion des métadonnées d'articles scientifiques disponibles en accès ouvert sur Internet* » (6, Nawrocki).

Rapidement, des institutions du domaine patrimonial, en particulier de grandes bibliothèques, ont repris le protocole OAI-PMH à leur compte pour en faire un outil permettant, par l'intermédiaire de la collecte de métadonnées, de proposer des **recherches transversales** (recherche à partir d'un sujet, d'un auteur sur des collections d'origines diverses) et **multi-sectorielles** (recherche dans des bibliothèques et des musées par exemple) et de faire connaître leurs ressources sur le web. Mais avant de développer ce point, définissons l'expression « domaine patrimonial » ; appartient au **domaine patrimonial** toute institution ayant pour mission de conserver, préserver, organiser, mettre en valeur et diffuser le patrimoine culturel. Ces organismes peuvent être des bibliothèques, des musées, des centres d'archives, des centres de documentation, etc…, publics ou privés. Les ressources qui constituent le patrimoine sont variées : livres, tableaux, documents sonores, cartes, objets, sites web, etc…

Ces dernières années, deux révolutions technologiques ont bousculé les institutions culturelles. Tout d'abord, le développement massif de l'**informatique**, avec les possibilités offertes en termes de développement et d'utilisation, a mené les acteurs du domaine patrimonial à informatiser leurs catalogues et à débuter un travail de numérisation de leurs collections pour des raisons de conservation mais également de diffusion. Puis, la forte expansion du **web** les a conduit à opérer une remise en question et à tenter d'occuper ce nouvel espace.

Ainsi, le domaine patrimonial a développé en interne des **bases de données***, reflet de leurs fonds, dont l'information finement structurée permet d'effectuer des recherches souvent très abouties dans leurs catalogues. A chaque ressource d'un fonds correspondent des métadonnées qui servent à sa **description** (par exemple, les champs *titre*, *auteur*, *éditeur*) et qui permettent de la retrouver. Les métadonnées constituent donc des moyens d'accéder à des fonds et de retrouver des ressources. Cependant, le traitement fin réservé à la description des ressources patrimoniales n'est pas mis en valeur sur la toile : les moteurs de recherche type Google n'indexent pas le **web dynamique**, c'est-à-dire les bases de données ; seules les pages dites statiques sont « lues » par ces moteurs. Ces bases de

données peu visibles constituent ce que l'on appelle communément aujourd'hui le « **web invisible*** » ou encore « web caché ».

D'aucuns ont donc vu dans l'émergence de l'OAI-PMH une possibilité de contourner ce problème de visibilité et d'offrir un **accès intégré** à un très grand nombre de ressources culturelles disséminées à travers le monde « *sans les dupliquer ni modifier leur localisation d'origine »* (6, Nawrocki).

Ce sont ces accès intégrés, ces **portails*** web, qui sont au cœur de ce mémoire.

Ce qui frappe le plus lorsqu'on lit la littérature sur le sujet, c'est l'idée de simplicité de mise en œuvre, qui revient comme un leitmotiv pour présenter l'OAI-PMH. Le protocole est en effet simple à comprendre car les concepts qui le gouvernent sont peu nombreux et la logique d'articulation est plutôt évidente.

Mais si l'on réfléchit aux objectifs de tels portails et que l'on interroge les moyens à utiliser pour les atteindre, la réalité tend à se complexifier. Le portail web offre un accès unifié à des ressources de nature et de provenance diverses. Il s'agit donc de créer de l'homogène à partir de sources hétérogènes. Pour retrouver ces ressources, il faut mettre en place des **fonctionnalités de recherche** sur le portail. Par nature, l'OAI-PMH permet une interopérabilité technique* simple vu qu'il détermine un langage et une syntaxe s'appuyant eux-mêmes sur des protocoles existants mais il s'arrête au **transfert de données**, il ne propose rien en termes de recherche.

C'est ce point-là que veut interroger ce mémoire ; comment la recherche, si essentielle et travaillée dans les applications locales des acteurs du domaine patrimonial, s'organise-t-elle au sein de ces nouveaux portails ? La quête de visibilité sur le web se fait-elle au détriment des fonctionnalités de recherche ? Les institutions patrimoniales perdent-elles en termes de recherche ce qu'elles semblent gagner en termes de diffusion ?

Après une première partie permettant de définir ce qu'est l'OAI-PMH et les raisons qui font que le domaine patrimonial se montre si prompt à l'adopter, une étude sur les fonctionnalités de recherche réalisée majoritairement à partir d'observations faites sur les métadonnées de quatre portails culturels permettra de mettre en lumière ce que l'OAI-PMH ne résout pas. La dernière partie sera consacrée aux préconisations dans la mise en œuvre du protocole OAI-PMH.

Première partie
L'OAI-PMH et le domaine patrimonial, les raisons d'un essor

1 L'OAI-PMH

1.1 Définition

L'OAI-PMH (Open Archives Initiative's Protocol for Metadata Harvesting ou Protocole de collecte de métadonnées de l'Initiative des Archives Ouvertes) est un protocole* permettant le partage de métadonnées*. Il définit un **standard** pour « *faciliter l'interopérabilité* de ressources documentaires, sans duplication ni déchargement des documents numériques primaires* ». (6, Nawrocki)

Mais avant d'aborder les rôles que définit le protocole et de détailler les aspects techniques essentiels, il est intéressant d'en rappeler l'historique.

1.2 Historique

L'OAI-PMH est né dans le cadre de l'Open Archive Initiative (Initiative des Archives Ouvertes) et fait partie de la Convention de Santa Fé de 1999. L'objectif était d'améliorer la circulation et la diffusion de l'information scientifique en permettant l'interopérabilité des différents systèmes d'auto-archivage des prépublications scientifiques des universités grâce à la définition d'un certain nombre de standards. Dans ce contexte, le terme « archive » est employé pour les prépublications scientifiques, c'est-à-dire les différentes versions d'un article avant acceptation de la version définitive par un comité de rédaction ou de lecture.

L'OAI-PMH est rapidement apparu intéressant pour l'échange de métadonnées et la création de nouveaux services dans des domaines autres que les sciences.

1.3 Concepts documentaires

Voici les concepts documentaires, suivis de leur équivalent en anglais, sur lesquels le protocole s'appuie (6, Nawrocki) :

- La **ressource** (*resource*) : c'est le document-objet (livre, table, image numérique, site web,...) qui est décrit ;
- L'**item** (*item*) : c'est la fiche ou la notice informatique qui décrit la ressource. L'item doit contenir un identifiant unique pour être compatible OAI-PMH ;
- L'**enregistrement** (*record*) : c'est l'ensemble des métadonnées extraites d'un item au format XML*. Il y a autant d'enregistrements par item que de formats

documentaires* dans lesquels est décrite une ressource (par exemple, une même ressource peut être décrite au format UNIMARC et Dublin Core) ;

- Le **lot** (*set*) : les items peuvent relever d'un ou plusieurs lots* pour permettre une moisson en bloc par type de support (périodique, image,...), par thème,...

Voici un schéma réalisé d'après des schémas de François Nawrocki et de Muriel Foulonneau (4, Foulonneau ; 6, Nawrocki) :

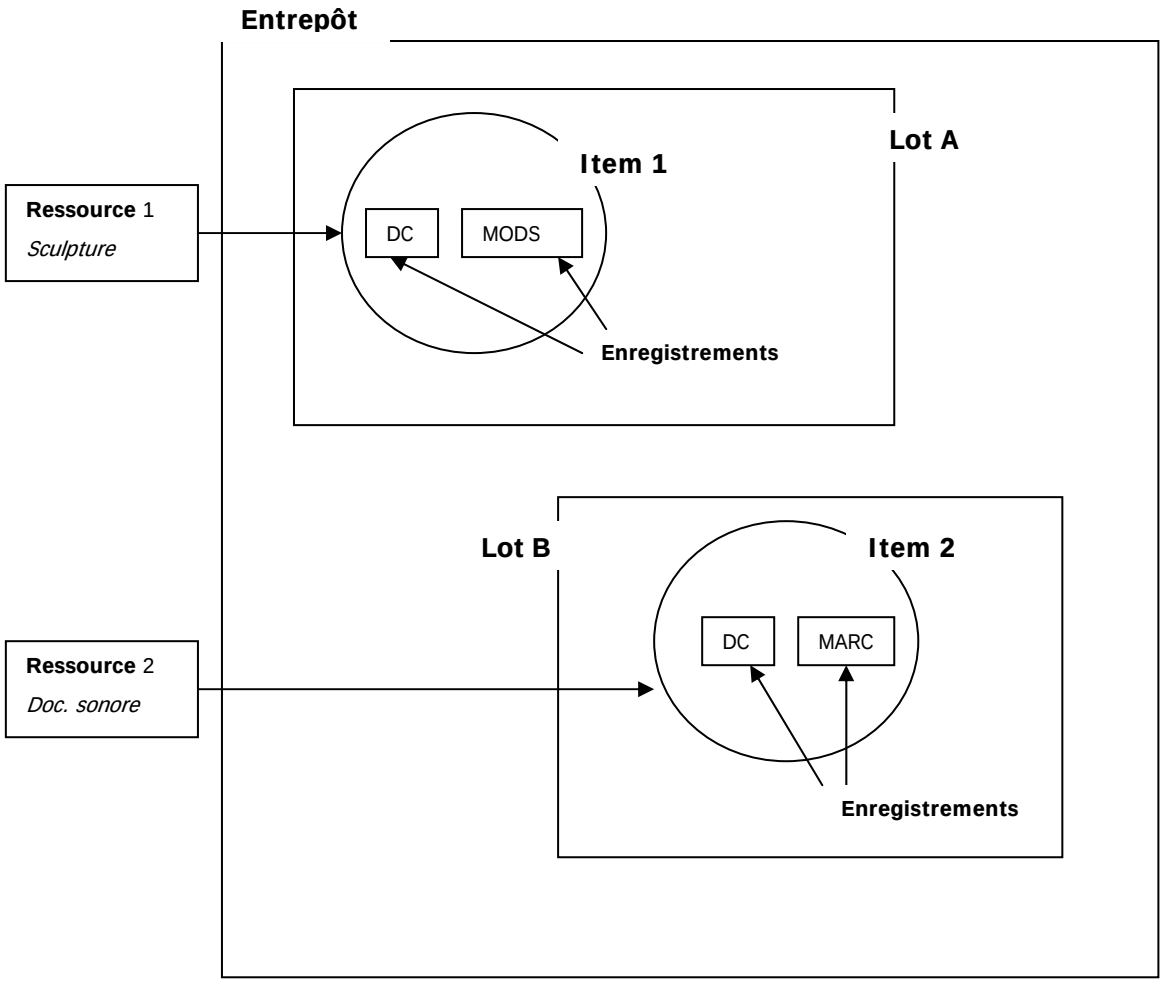


Fig. 1 : Concepts documentaires de l'OAI-PMH

Le protocole définit les rôles des différents acteurs dans l'échange de leurs métadonnées.

1.4 Principes organisationnels

L'OAI-PMH met en place un système d'échange de métadonnées en **architecture distribuée*** avec pour acteurs des **fournisseurs de données*** (*data providers*), des **agrégateurs*** (*aggregators*) et des **fournisseurs de service*** (*service providers*) qui moissonnent les métadonnées des fournisseurs de données rapatriées dans un **entrepôt***. On dit de l'architecture qu'elle est distribuée car les ressources dont on collecte les métadonnées sont réparties dans différentes bases (5, Gatenby).

1.4.1 Le fournisseur de données

Le fournisseur de données (FD) : il prépare, met à disposition et assure la maintenance des métadonnées de son choix dans un entrepôt OAI. Bien qu'il puisse proposer ses métadonnées en autant de formats qu'il le désire, le format Dublin Core non qualifié est requis par le protocole pour assurer une interopérabilité minimale. Le FD peut choisir de mettre à disposition ses données selon un modèle partageable* (tout fournisseur de services ou agrégateur a accès à l'entrepôt) ou réservé* (l'accès est limité à des fournisseurs de services choisis).

1.4.2 Le fournisseur de services

Le fournisseur de services (FS) : il propose via une interface unique l'exploitation des métadonnées qu'il collecte de fournisseurs de données grâce à un programme appelé moissonneur*. Cette exploitation se traduit souvent par une interface de recherche plus ou moins élaborée et des services additionnels tels que la fonction panier, l'espace utilisateur personnalisé, etc...

1.4.3 L'agrégateur

L'agrégateur : c'est un intermédiaire optionnel entre le FD et le FS. En effet, il arrive que les FD ne puissent maintenir un entrepôt OAI. L'agrégateur a dans ce cas-là pour fonction de rassembler les métadonnées de différents FD, de les retraiter pour les normaliser et de les mettre à disposition dans un entrepôt OAI.

Le protocole OAI-PMH permet le partage de métadonnées et non de ressources. Les ressources restent localisées chez le FD. En outre, il ne gère que le **transfert des données** ; il ne propose aucune fonction de recherche. C'est au FS de mettre en place des **applications** permettant la recherche de notices à partir de l'entrepôt.

Voici, pour résumer le fonctionnement de l'OAI-PMH, la reproduction d'un schéma de François Nawrocki (6, Nawrocki) :

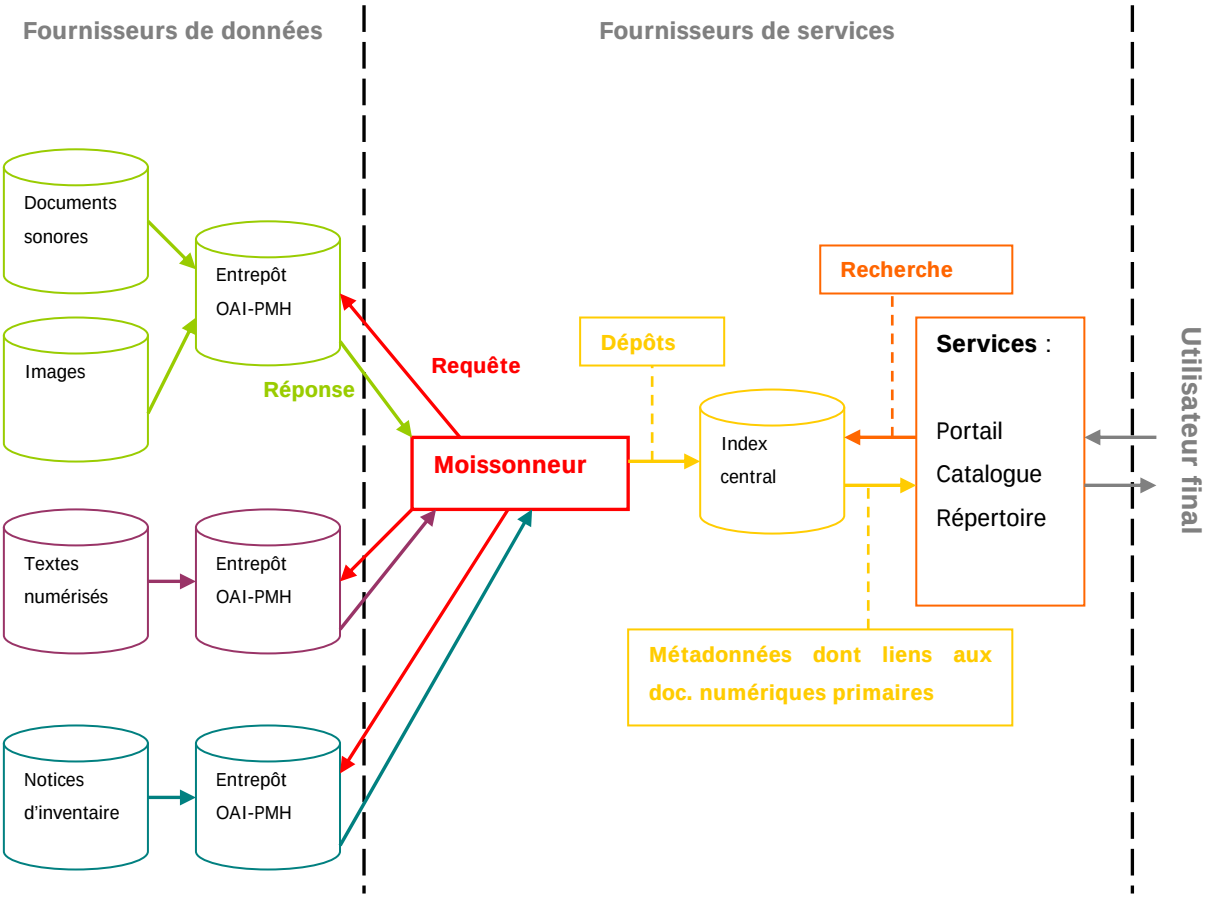


Fig. 2 : Organisation fonctionnelle en OAI-PMH

1.5 Aspects techniques

Le FS, par l'intermédiaire du moissonneur, soumet des requêtes XML* aux entrepôts en recourant aux protocoles HTTP* et URL*. Les réponses sont également en syntaxe XML.

1.5.1 L'entrepôt

Le FD crée un entrepôt OAI dans lequel il met à disposition un ensemble d'enregistrements qui peuvent se matérialiser selon différents formats de description. Il revient donc au FD de sélectionner les données qu'il veut rendre moissonnables.

Chaque entrepôt peut être divisé et sub-divisé en ensembles ou lots de données présentant des critères communs. Ainsi, les lots sont souvent organisés selon le sujet ou le type de document (4, Foulonneau).

C'est cet entrepôt qui va recevoir les requêtes d'un programme appelé moissonneur.

1.5.2 Le moissonneur

Le moissonneur est un programme lancé par le FS pour interroger des entrepôts OAI dont il collecte les données qui l'intéressent. Le protocole a défini six requêtes simples qui lui permettent de récupérer des enregistrements et de les mettre à disposition dans le réservoir ou index central qui sera consulté directement par l'utilisateur final (4, Foulonneau ; 6, Nawrocki).

1.5.3 Les requêtes

Les requêtes sont au nombre de six (4, Foulonneau ; 6, Nawrocki) :

- **Identify** : information sur un entrepôt OAI
- **ListMetadataFormats** : liste des formats disponibles dans l'entrepôt
- **ListSets** : liste des lots formés par les FD
- **ListIdentifiers** : liste des identifiants et date de dernière modification des items disponibles
- **GetRecord** : demande d'un enregistrement précis
- **ListRecords** : récupérer en bloc l'ensemble des items d'un répertoire

Toutes les métadonnées ne sont pas collectées à chaque passage du moissonneur : seuls les nouveaux enregistrements, les modifications et les suppressions d'enregistrements sont pris en compte. On dit alors que le modèle est **asynchrone** car la mise à jour de l'entrepôt et la collecte de ces mises à jour ne sont pas simultanées (4, Foulonneau).

1.6 Aspect documentaire, le format Dublin Core

Afin d'assurer une interopérabilité technique minimum, en plus de l'encodage des données en XML et de l'utilisation du langage HTTP, les FD doivent mettre à disposition leurs données sous le format **Dublin Core non qualifié** (parfois appelé Dublin Core simple) au minimum.

1.6.1 Historique et objectifs

Le format **Dublin Core** (DC) a été créé en 1995 à l'instigation de professionnels de disciplines diverses telles que la bibliothéconomie, l'informatique, le balisage de textes, la communauté muséologique et autres domaines connexes afin de répondre aux objectifs suivants (10, Hillmann) : « *proposer un format de description simple à créer et à gérer afin de faciliter la recherche transversale par domaine sur le web* » (11, Lagoze). En effet, dans un contexte de profusion d'information tel qu'on le trouve sur la toile, se fait jour un besoin d'organiser les informations en leur appliquant des métadonnées minimum simples.

Parallèlement à cet objectif de simplicité, toujours d'actualité, s'est développée une demande de la part de communautés et de domaines spécifiques d'un format permettant une description plus fine, demande qui s'est matérialisée par la création de **Dublin Core qualifié** *(voir 1.6.2.2)*.

1.6.2 Caractéristiques

Le format Dublin Core se veut un format de description simple, une sorte d'esperanto que diverses communautés pourraient comprendre, intégrer et utiliser pour un minimum d'interopérabilité. Trois caractéristiques principales définissent le format DC (10, Hillmann) :

- **Simplicité** du format : il n'existe que **15 éléments** ou champs pour décrire une ressource ;
- **Sémantique** communément comprise : les éléments représentent des concepts censés être universellement compris (les éléments *créateur, titre*) ;
- **Extensibilité** : réalisée grâce aux qualificatifs DC (*voir 1.6.3.2 Dublin Core qualifié)*, elle permet d'affiner la description de la ressource.

1.6.2.1 Dublin Core non qualifié ou simple

Le format Dublin Core non qualifié se compose de 15 éléments de base (le *core* de *Dublin Core* en anglais). On peut les diviser en 3 grandes catégories (13) :

Contenu	**Propriété**	**Instanciation**
Couverture	Collaborateur	Date
Description	Créateur	Format
Relation	Droits	Identifiant
Source	Editeur	Langue
Sujet		
Titre		
Type		

Tab. 1 : Les éléments Dublin Core

1.6.2.2 Dublin Core qualifié

En juillet 2000, la DCMI (Dublin Core Metadata Initiative, l'organisation qui travaille sur le format DC) a émis une liste de **qualificatifs** dans le but d'affiner la description des ressources sous le format DC. Ils sont classés en 2 grandes catégories (10, Hillmann) :

- **Raffinement d'éléments** (*element refinement*) : ces qualificatifs permettent de préciser un élément. Par exemple, l'élément *Créateur* peut être précisé par *Parolier* ou par *Compositeur*. L'élément qualifié devient un spécifique de l'élément de base mais il reste de même nature ;

- **Schéma d'encodage** (*value encoding*) : ces qualificatifs permettent d'identifier le contexte de la valeur d'un élément. Les schémas peuvent être des vocabulaires contrôlés, des normes. Par exemple, l'élément Sujet de valeur 190 (Philosophie occidentale) peut être complété par le qualificatif Dewey pour signifier que la classification Dewey est utilisée pour indexer les ressources décrites. Un autre exemple : l'élément Date peut avoir pour qualificatif ISO8601 pour spécifier que la date encodée sous la forme 2007-08-23 respecte la norme ISO8601.

Nous verrons par la suite que cette simplicité pose divers problèmes d'interprétation - malgré une description et une définition des éléments sur le site DC, il existe des ambiguïtés - et de mise en œuvre - certains FD créent des qualificatifs trop ancrés dans leurs pratiques, limitant ainsi l'interopérabilité technique et sémantique. Par ailleurs, certains types de ressources du domaine culturel, décrits dans des formats riches, peuvent perdre sémantiquement au passage en format DC.

1.7 Les droits

Il faut bien distinguer l'expression des droits relatifs aux **métadonnées** et à la **ressource**.

Dans le cadre de l'OAI-PMH, l'expression des droits concernant les métadonnées peut se faire à trois niveaux :

- Au niveau de l'**enregistrement** : la spécification *rights* peut être optionnellement encapsulée dans la balise *<about>* d'un enregistrement ;
- Au niveau de l'**entrepôt** : la balise *<description>* peut contenir optionnellement des renseignements sur les types de droits applicables aux enregistrements de l'entrepôt ;
- Au niveau du **lot** : ici, c'est la balise *<setDescription>* qui permet la mention de droits.

Ces indications sont utiles aux FS qui vont utiliser les métadonnées. Par exemple, un FD peut indiquer dans ces balises qu'il refuse toute modification de ses métadonnées (4, Foulonneau).

La mention des droits relatifs à la ressource est également possible (élément *DC_rights*) ; elle sert alors à indiquer à l'utilisateur quels droits sont applicables à la ressource numérique. Cet élément peut être précisé par les qualificatifs *accessRights* (qui peut avoir accès à la ressource) et *License* (document légal fixant officiellement les conditions d'utilisation d'une ressource) (10, Hillmann ; 13).

2 L'OAI-PMH dans le domaine patrimonial

2.1 Une utilisation « détournée »

Comme nous l'avons vu précédemment, le protocole OAI-PMH a été originellement mis en place dans le domaine des prépublications scientifiques. Il visait à décrire de façon simplifiée grâce au format Dublin Core des documents académiques, donc tous de même nature. Rapidement, la Bibliothèque du Congrès (l'équivalent de notre BNF aux Etats-Unis) et les bibliothèques de grandes universités américaines se sont saisies de ce protocole et l'ont utilisé dans le but de faire découvrir à un public le plus large possible les ressources qu'elles possédaient dans leurs bases de données.

D'un protocole permettant de connaître la description de travaux scientifiques en cours, l'OAI-PMH est devenu un outil assurant la centralisation de métadonnées hétérogènes dans leur nature et leur provenance afin offrir un **accès simplifié à des ressources numériques diverses et disséminées**.

2.2 Raisons d'un fort développement de projets s'appuyant sur l'OAI-PMH

Pour le domaine patrimonial, l'OAI-PMH paraît être en mesure de répondre à des problématiques de **valorisation** et de **diffusion** immémoriales. Ces institutions semblent avoir trouvé dans ce protocole une nouvelle façon de renforcer leur mission de valorisation et de diffusion du patrimoine acquis, traité intellectuellement et conservé.

En outre, les portails agrégés permettent aux internautes d'accéder plus facilement à une **pluralité de ressources** au niveau des domaines traités et des types. Ils sont particulièrement intéressants lorsqu'on veut réaliser une **recherche transversale** (par exemple, recherche à partir d'un mot-clé ou d'un type de ressources sur des collections différentes). En outre, ils évitent à l'utilisateur d'effectuer une même recherche à des endroits différents.

Voici les principales raisons couramment invoquées par le domaine patrimonial pour développer des projets en OAI-PMH.

2.2.1 Guichet unique

La création d'interfaces uniques permettant une recherche dans des bases diverses présente pour l'utilisateur un intérêt pratique indéniable. Muriel Foulonneau (4, Foulonneau) résume

bien cette idée d'accès simplifié aux ressources du domaine patrimonial : *« Face à la fragmentation des connaissances culturelles en ligne, de nouveaux services peuvent fournir un accès clair et cohérent au capital culturel numérique ».*

2.2.2 Visibilité et élargissement des publics

Les acteurs du domaine patrimonial ont commencé à entreprendre ces dernières années pour des raisons liées à leurs missions de conservation mais aussi de diffusion, de vastes programmes de numérisation de leurs fonds. Par ailleurs, l'émergence de l'informatique et du web a favorisé le développement de documents qui n'existent que sous forme numérique (*born-digital documents* en anglais), tel qu'un site web. Internet est devenu alors pour eux un outil permettant de faire découvrir leurs ressources.

Cependant, les moteurs de recherche type Google n'indexent pour le moment que le web statique. Ainsi, toutes les informations structurées contenues dans les bases de données accessibles en ligne gérées par les institutions culturelles (contenu des sites dynamiques) ne sont pas identifiées par ces moteurs. Le protocole OAI-PMH permet aux organismes le désirant de donner accès à leurs données par l'intermédiaire de portails et d'accroître ainsi leur **visibilité** sur le web (2, Arms ; 3, Boston ; 4, Foulonneau ; 5, Gatenby ; 7, Prom).

Le protocole permet également la **multiplication des points d'entrée** aux ressources d'un organisme vu qu'il est possible de rendre ses métadonnées moissonnables par plusieurs fournisseurs de services.

Cette nouvelle visibilité sur l'espace en développement exponentiel qu'est le web peut également contribuer à élargir le public susceptible d'être intéressé par un accès aux ressources culturelles numériques (2, Arms ; 4, Foulonneau).

2.2.3 Transversalité

Le protocole, en proposant une interopérabilité technique* simplifiée, permet la centralisation de données de nature et de provenance diverses. A charge pour le FS de proposer des services autour de ces données (recherche croisée sur des fonds provenant de musées et de bibliothèques par exemple). Mais l'**interopérabilité technique** permise par le protocole OAI-PMH n'induit pas forcément une **interopérabilité sémantique*** et **organisationnelle***, comme nous le verrons par la suite. En effet, les passerelles requises (type vocabulaire contrôlé pour l'indexation de documents) pour effectuer une **recherche transversale efficace** sur des bases ou des collections différentes ne sont pas

préexistantes dans le protocole OAI-PMH : elles sont à mettre en place par les acteurs (4, Foulonneau).

2.2.4 Facilité d'implémentation apparente

Les principes organisationnels de l'OAI-PMH sont assez simples à comprendre, ce qui amène de nombreux acteurs du domaine culturel à penser que le protocole est également simple à mettre en œuvre, surtout pour les fournisseurs de données, qui « n'ont qu'à » mettre à disposition dans un entrepôt leurs métadonnées au format Dublin Core non qualifié au minimum.

Cependant, de nombreuses limites à cette idée répandue doivent être émises (et seront développées par la suite). Tout d'abord, la plupart des institutions culturelles décrivent leurs ressources dans un format riche. Créer des équivalences entre un format riche (par exemple UNIMARC) et un format simple ou appauvri (Dublin Core et ses 15 éléments) pose des questions évidentes qui touchent à la qualité des métadonnées. Ensuite, les institutions dont les moyens sont limités financièrement et par là même techniquement auront plus de difficultés à créer leur entrepôt et à en assurer la maintenance (21, Shreeves). En effet, la mise en place d'un entrepôt OAI fait intervenir tout un panel de compétences relatives aux (4, Foulonneau) :

- **Technologies de l'information** (langage XML, scripts pour l'extraction des métadonnées du système source à encoder en XML, réseaux) ;
- **Métadonnées** (connaissance des fonds exposés, règles de catalogage, sélection de l'information) ;
- **Stratégie de la diffusion des données** (question des droits, accès des utilisateurs)

Quant aux fournisseurs de services, leur tâche de développement de fonctionnalités de recherche efficaces à partir de données disparates est plus ardue qu'il n'y paraît, comme nous le verrons dans la deuxième partie du mémoire (8, Shreeves).

2.2.5 Contrôle sur les ressources

Les FD gardent le contrôle sur leurs ressources : ils mettent à disposition les métadonnées qu'ils veulent dans des entrepôts ouverts à tout moissonneur ou réservés. Les documents primaires restent localisés chez eux ; seules les métadonnées sont regroupées dans l'entrepôt et l'index central ou réservoir. Ils continuent également de gérer les **droits** de diffusion et d'auteur associés à leur fonds (4, Foulonneau ; 6, Nawrocki).

2.2.6 Coûts

L'investissement technique et financier est limité pour le FD, bien que l'on puisse émettre les mêmes objections que précédemment *(voir 2.2.3 Facilité d'implémentation relative)*. En revanche, ce qui est parfois vrai pour le FD ne l'est pas pour le FS qui a pour charge, à partir de données hétérogènes, de fournir une interface unique facilement compréhensible par les utilisateurs finaux et de proposer des fonctionnalités de recherche pointues. En résumé, le FS doit proposer de l'homogène à partir de données hétérogènes, ce qui va se révéler, comme nous le verrons grâce à l'étude en deuxième partie, souvent problématique (8, Shreeves).

2.3 Conséquences organisationnelles

2.3.1 Nouveaux partenariats

La simplicité technique apparente de l'OAI-PMH permet de développer de nouveaux types de **partenariats** et de **projets** entre des communautés de mêmes domaines mais aussi de domaines connexes.

Ainsi, selon Shreeves (21, Shreeves), le monde des bibliothèques numériques utilise le protocole pour atteindre divers objectifs :

- **Accès intégré à des ressources disséminées** : avec l'exemple du site OAIster, dont les fonctionnalités de recherche seront étudiées en deuxième partie ;
- **Regroupement de ressources de même format** : le site du Sheet Music Consortium, qui réunit les catalogues d'universités américaines autour de la musique en feuilles numérisée (également étudié en deuxième partie) ;
- **Regroupement de services spécifiques adaptés à un public particulier** : exemple de la National Sciences Digital Library[1] (NSDL) qui donne accès à des ressources pédagogiques en sciences à destination de professeurs et d'élèves ;
- **Regroupement de ressources diverses autour d'une même thématique** : le site Aquitaine patrimoines[2] (BNSA - Banque Numérique du Savoir d'Aquitaine), qui donne accès à des ressources numérisées provenant de diverses institutions patrimoniales de la région Aquitaine telles que services d'archives, centres de documentation, musées.

[1] Site de la NSDL : http://nsdl.org/
[2] Site de la BNSA : http://bnsa.patrimoines.aquitaine.fr

2.3.2 Nouvelles communautés de pratiques

Ces projets et partenariats émergents, qui amènent des professionnels de différents horizons à travailler ensemble, créent de nouvelles **communautés de pratiques***. L'interopérabilité technique étant nécessaire mais souvent non suffisante, les parties d'un projet et les professionnels d'un domaine doivent s'entendre et travailler ensemble sur des normes et standards afin d'assurer une interopérabilité sémantique et syntaxique. Il apparaît également nécessaire d'effectuer un **retour d'expérience** et de rédiger des **guides de bonnes pratiques** afin de capitaliser les connaissances acquises dans le domaine. Comme nous le verrons en troisième partie, la production de bonnes pratiques est importante, en particulier aux Etats-Unis.

Le développement de portails selon le standard OAI-PMH trouve également sa place dans la notion émergente de **web 2.0***. En effet, le web regorge d'outils permettant le **travail collaboratif** et la **mutualisation** des connaissances sur le sujet tels que wikis de bonnes pratiques, tutoriels sur le protocole, publication de tableaux d'équivalence entre différents formats de description, forums de discussion, blogs, etc…

L'OAI-PMH, associé au format de description Dublin Core, est né d'une volonté de simplifier les échanges de métadonnées et par là même l'accès à des ressources. La relative facilité de mise à disposition des données dans des entrepôts ne doit cependant pas cacher les difficultés posées dans la mise en œuvre par les FS de services de qualité, notamment au niveau des fonctionnalités de recherche. En effet, toute la difficulté consiste à rendre homogènes des notices très disparates (au niveau des institutions, des formats, des règles d'écriture), l'interopérabilité technique permise par l'utilisation de protocoles ne faisant pas forcément l'interopérabilité sémantique, syntaxique et organisationnelle.

Une étude de portails utilisant le protocole OAI-PMH va permettre de faire ressortir un certain nombre des difficultés que rencontrent FD et FS dans la création de portails intégrés à la recherche efficace.

Deuxième partie
Etude des fonctionnalités de recherche de portails utilisant le protocole OAI-PMH

1 Contexte de l'étude : le projet de création d'un portail sur la musique contemporaine

La présente étude a été réalisée à l'occasion d'un stage à la documentation de la Cité de la Musique. Elle a pour ambition de dresser un panorama de la recherche proposée sur des « portails OAI-PMH » agrégeant des notices provenant d'institutions diverses.

La Cité de la Musique est partenaire d'un projet de création pour fin 2007 d'un portail OAI-PMH sur la musique contemporaine avec d'autres médiathèques spécialisées françaises. Jusqu'à présent, deux difficultés dans la mise en œuvre de ce projet ont été identifiées : le traitement des notices d'autorité et le chemin d'accès au document numérique, points qui ont guidé la présente étude. En outre, certaines institutions ont des difficultés à isoler leurs notices relatives à la musique contemporaine, contrairement à la Cité de la Musique qui a pu réaliser une sélection parmi son fonds grâce au descripteur « Musique contemporaine » de son thésaurus qui a joué le rôle de critère de dissociation.

Avant d'aborder les résultats de l'évaluation, voici une présentation[1] du projet et de ses acteurs.

1.1 Acteurs

Ils sont tous spécialisés dans le domaine de la musique ou de la musique contemporaine et sont au nombre de six.

- **La Cité de la Musique**

  La Médiathèque propose un fonds constitué de livres, périodiques, plans d'instruments, documents sonores et vidéos principalement sur la musique savante, le jazz, les musiques du monde et l'organologie.

  Elle proposera un fonds sur la Musique contemporaine d'environ 8300 notices.

- **L'IRCAM (Institut de Recherche et Coordination Acoustique/Musique)**

  L'IRCAM est un institut de recherche dédié à la musique contemporaine. L'institut fournira environ 36400 notices en Musique contemporaine.

[1] Partie réalisée grâce à la présentation suivante : FINGERHUT Michel. Un portail pour la musique contemporaine, juin 2007. Ircam, Ministère de l'Education Nationale. 11 p.

L'IRCAM, à l'instigation d'un projet qui ne consistait au départ qu'à proposer un accès unifié à ses ressources propres, assure la direction technique.

- **Le Centre de Documentation de la Musique Contemporaine (CDMC)**

  Ce centre permet la consultation d'œuvres contemporaines (partitions, enregistrements, documentation sur les auteurs).

  Nommé institution pilote, il proposera au sein du portail environ 49200 notices relatives à la musique contemporaine.

- **Le Conservatoire National Supérieur de Musique de Paris (CNSMDP)**

  Le catalogue couvre l'ensemble de l'histoire de la musique et de la danse. Le Conservatoire fournira environ 33500 notices en musique contemporaine.

- **L'Ensemble Inter-Contemporain (EIC)**

  Il a pour mission de diffuser, transmettre et créer la musique contemporaine du 20ème siècle à aujourd'hui.

  Sa contribution au projet se fait sous la forme d'environ 790 heures de documents sonores exclusivement.

- **La Médiathèque Musicale Mahler**

  La médiathèque possède un fonds spécialisé sur la musique classique du Moyen-Age à nos jours. Elle contribuera à hauteur d'environ 10200 notices.

1.2 Objectifs du projet et public-cible

C'est un projet hybride, financé dans le cadre du plan de numérisation de la MRT (Mission Recherche et Technologie) du Ministère de la Culture : il s'agit d'effectuer un travail de **numérisation des collections** pour des questions de conservation mais également d'offrir un **accès unique** à des ressources disséminées.

Ainsi, d'après une note du CDMC[1],

> *Ce portail a pour ambition de fédérer l'accès en réseau aux ressources concernant la création musicale contemporaine et ses acteurs, gérées par les organismes*

[1] Appel à projets MRT 2007 : le portail internet de la musique contemporaine, septembre 2006. Centre de Documentation de la Musique Contemporaine. 3 p.

spécialisés.

L'articulation de la numérisation et de l'accès fédéré à ces fonds servira à accroître la visibilité de ce domaine et de ses acteurs.

Le portail, dont le domaine est plutôt pointu, s'adresse à des professionnels de la musique et à un public amateur averti.

Ainsi, élèves et étudiants, professionnels du domaine et de domaines connexes (tels que la danse) et mélomanes sont les destinataires de ce projet.

1.3 Contenus

Une recherche centralisée permettra l'identification, la localisation et l'accès – à distance ou localement – aux documents, ressources et informations concernant ce domaine.

Toujours d'après la même note du CDMC,

> *Le portail fournira les réponses correspondantes jusqu'au niveau le plus détaillé (ouvrage, partition, enregistrement, biographie, événement…) et permettra d'y accéder – sur l'internet, dans un extranet de partenaires, dans l'intranet ou dans les murs de l'organisme dépositaire, en fonction des droits.*

Les contenus, bien qu'appartenant au même thème, sont de nature très diverse.

- **Typologie des ressources**

  Les ressources sont en priorité des ressources numériques. On trouvera ainsi sur le portail Musique contemporaine des :

  - Documents (lettre, livre, partition, enregistrement sonore, vidéo, note de programme) ;
  - Evénements (concert, formation) ;
  - Personnes (compositeur, musicien) ;
  - Organismes.

- **Format des ressources**

  Les ressources seront décrites en Dublin Core et en MODS (Metadata Object Description Schema, de la Bibliothèque du Congrès) qui est un format s'appuyant sur MARC21 et permettant de le réutiliser.

- **Accès aux ressources**

  Sous réserve d'accord, les ressources numériques sonores et audiovisuelles seront consultables dans leur intégralité depuis l'intranet des partenaires (donc sur place), sous forme d'extraits longs depuis l'extranet (chez chacun des partenaires) et sous forme d'extraits courts sur internet. Des droits de diffusion élevés et difficiles à négocier rendent la consultation sur le web des ressources dans leur intégralité impossible pour le moment.

2 Méthodologie

2.1 Pour la sélection de portails

La sélection des quatre portails retenus pour l'étude a été réalisée grâce à la consultation de ressources de nature diverse. Il a été décidé, en commun accord avec le directeur de stage, de travailler uniquement sur des portails du domaine culturel qui mènent à la **consultation de ressources numériques** et de limiter le nombre de portails étudiés.

2.1.1 Personnes-ressources

La rencontre de personnes du domaine patrimonial travaillant sur le projet Musique contemporaine mais aussi sur d'autres projets utilisant l'OAI-PMH a permis de faire une large pré-sélection de portails. Le coordinateur du catalogue et responsable de la documentation du musée et le responsable du service système d'information de la Cité de la Musique, tous deux mobilisés pour le projet Musique contemporaine, ont d'abord été consultés.

Au CDMC, un entretien avec une documentaliste/administratrice de la base de données également mobilisée sur le projet Musique contemporaine a permis de recueillir des informations complémentaires sur le sujet.

Une rencontre avec un responsable du Département de la Bibliothèque Numérique de la BNF a permis d'élargir le nombre de portails éligibles mais également de confirmer la pré-sélection de portails.

Les portails sélectionnés appartiennent donc tous au domaine patrimonial. En outre, la fiabilité de l'information proposée est garantie par les personnes consultées et par des critères d'évaluation déterminés dans la grille mise en place.

2.1.2 Grille de sélection

Cf annexe 2 p. 109

La constitution de la grille présidant à la sélection des portails a été réalisée d'après des travaux sur l'évaluation de ressources du web. D'après Alexander et Tate de l'université de Widener (14, Alexander), cinq critères généraux permettent de juger de la fiabilité d'un site web. En voici quatre qui ont été utiles dans la constitution de la grille. En premier lieu, l'autorité doit être clairement identifiable (nom, adresse pour le contact, qualité de l'auteur). L'information doit être précise, c'est-à-dire que les sources doivent être clairement citées et

que le site doit être exempt de fautes d'orthographe ou d'approximations syntaxiques. L'objectivité est le troisième critère et peut se traduire par la question suivante : l'information fournie est-elle mêlée à de la publicité ? Le troisième critère concerne les dates (date de création d'une page/d'un site, date de publication, dates de mise à jour et toute information concernant les procédures de mise à jour).

Alexandre Serres, de l'URFIST de Rennes (16, Serres) propose également une grille d'identification et d'évaluation de sites web elle-même divisée en cinq sous-grilles (qui ne seront pas développés dans leur ensemble). L'organisation responsable du site doit être clairement et facilement identifiable ; son sérieux, sa compétence et sa notoriété dans le domaine doivent être avérés, et les dates de création et de mise à jour du site doivent être clairement indiquées. L'évaluation du contenu permet également de juger de la fiabilité d'un site web. Cette sous-grille comporte des critères de degré de fiabilité, de niveau de précision, de degré de nouveauté, de fraîcheur de l'information, de clarté dans l'indication des ressources, de pertinence et richesse des liens externes, de qualité de la langue et de clarté de présentation.

2.2 Pour l'évaluation de la recherche

Cf annexe 1 p. 108

Le but de l'étude étant d'évaluer la qualité de la recherche au sein de portails dont la particularité est d'utiliser le protocole OAI-PMH, une modélisation d'une recherche en ligne s'imposait pour pouvoir en identifier les différentes étapes et déterminer des critères pour évaluer ces étapes. La modélisation d'une recherche en ligne s'est appuyée en partie sur le travail réalisé par Pascal Duplessis (15, Duplessis).

Ainsi, de cette modélisation a découlé la création d'une grille (*cf annexe 3 p. 110*) mettant tout particulièrement l'accent sur les points d'accès, le traitement des autorités et des vocabulaires contrôlés, les formats d'affichage des notices, les valeurs des champs et le chemin d'accès au document numérique, point particulièrement important pour le projet Musique contemporaine.

La grille évalue donc trois points principaux : le formulaire de **recherche simple**, le formulaire de **recherche avancée** et l'**affichage** des résultats. Une quatrième partie modélise le **chemin d'accès** au document numérique en numérotant le nombre de clics nécessaires pour l'atteindre.

Vu l'hétérogénéité des portails et des ressources, l'établissement d'une recherche-type systématique était impossible. Pour ce qui est d'une analyse systématique sur vastes

échantillons de notices et de leurs éléments, difficile à réaliser dans notre cadre, de nombreuses études à large échelle sur le sujet existent (*voir Bibliographie, Partie* Qualité des métadonnées *p. 95*). Le parti pris a été de focaliser l'analyse sur les concepts documentaires garants d'une interrogation performante et fine, ce qui a amené à effectuer, pour chaque point, un type de recherche :

Concept documentaire	**Points à analyser**	**Type de recherche à effectuer**
Normalisation (assure la transversalité)	Index et listes contrôlées	• Point d'accès *auteur* • Point d'accès *sujet* (*subject*, *keyword*) • Point d'accès *type*
	Règles d'écriture	Elément *date*
	Fonction tri	• Tri par *date* • Tri par *type*
Granularité	Relations entre les ressources décrites	• Noter le niveau de granularité décrit (niveau collection, objet, les deux ?) • Recherche d'une image appartenant à une collection • Repérer s'il existe l'élément DC *relation* dans les notices
Accessibilité / usabilité	Chemin pour accéder au document numérique	• Recherche sur le type ressource/document numérique • Nombre de clics à effectuer • Recherche de la présence de l'élément droits (*rights*)

Tab. 2 : Méthodologie de recherche

La notion de normalisation des métadonnées comme garante d'une recherche pointue, transversale et efficace sera largement développée en introduction de la troisième partie.

3 Présentation des portails étudiés

Les portails, au nombre de quatre, sont les suivants : **The European Library**[1] (TEL), **Europeana**[2], **Sheet Music Consortium**[3] et **OAIster**[4] (complété par Digital Library Federation). Il est à noter que la **Digital Library Federation**[5] (DLF) est un prototype développé par OAIster dans le but de tester de nouvelles fonctionnalités de recherche. Les ressources de la DLF sont donc composées d'une partie des ressources d'OAIster ; seuls les points d'accès changent.

Ils peuvent être divisés en deux catégories : les **portails spécialisés** ou pointus, qui offrent une sélection de ressources d'un domaine ou d'un type particulier tels que Sheet Music Consortium et Europeana et les **portails généralistes** ou encyclopédiques, qui ont pour ambition d'offrir un accès unifié à des ressources très diverses, tant dans leur type (document sonore, image, livre, texte,...) que dans leur provenance (bibliothèque, centre de documentation, archives,...) ou dans le sujet traité. OAIster (donc DLF) et TEL présentent tous deux ce profil. En d'autres termes, on peut dire que la première catégorie appartient à des projets bien **balisés** alors que la seconde regroupe des sortes de « méta-bibliothèques » recherchant l'**exhaustivité** dans leur registre (OAIster dans le registre académique et TEL dans le registre des bibliothèques nationales européennes).

La partie qui suit a pour but de présenter succinctement les portails en développant les points suivants : organisme responsable, type et nombre de notices, formats de description, publication en ligne de bonnes pratiques à l'attention des fournisseurs de données. Ce dernier point, qui sera l'objet de la troisième partie, permet de trouver des informations sur les formats documentaires et donne une bonne indication des efforts fait en termes d'interopérabilité sémantique, syntaxique et organisationnelle.

3.1 Portails spécialisés

- **Sheet Music Consortium**

  Mis en ligne en septembre 2003, le Sheet Music Consortium donne accès à environ 120 000 notices de musique en feuilles numérisée (partitions de quelques feuillets de musique et chansons populaires). Le producteur du site est l'UCLA (University of

[1] www.theeuropeanlibrary.org
[2] www.europeana.eu
[3] http://digital.library.ucla.edu/sheetmusic/
[4] www.oaister.org
[5] http://quod.lib.umich.edu/i/imls/

California, Los Angeles). Les universités d'Indiana, Johns Hopkins et Duke ainsi que la Bibliothèque du Congrès mettent à disposition leur fonds en la matière.

Le site donne accès à un guide des bonnes pratiques peu développé à l'attention des fournisseurs de données. Les éléments du format DC simple, utilisé sur le portail, y sont explicités ; un guide d'écriture et des conseils quant à l'utilisation de vocabulaires contrôlés (mais sans en nommer aucun) sont également présents.

- **Europeana**

Europeana est un prototype de bibliothèque en ligne préfigurant la Bibliothèque Numérique Européenne développée entre autre par la BNF. Le portail met à disposition un fonds encyclopédique d'environ 12 000 livres numérisés et libres de droit provenant de la BNF, de la Bibliothèque Nationale du Portugal (BNP) et de la Bibliothèque Nationale Széchényi de Hongrie. Le site ne propose pas de guide de bonnes pratiques.

3.2 Portails généralistes

- **The European Library**

Produit par la CENL (Conférence Européenne des directeurs de Bibliothèques Nationales), mis en ligne en mars 2005, The European Library (TEL) donne un accès intégré aux catalogues et aux collections numériques des bibliothèques nationales de 31 pays européens. 150 millions de ressources bibliographiques et numériques sont consultables à partir du portail. Les recherches effectuées pour les besoins de l'étude ne se feront que dans le catalogue des ressources numériques (onglet *Collections*, sélection *Matériaux numérisés*).

TEL édite en ligne un guide très riche listant conseils et bonnes pratiques à destination des fournisseurs de données désireux de rendre leurs ressources accessibles par le portail. Le format requis est, comme le stipule le protocole OAI-PMH, le Dublin Core. TEL conseille vivement d'utiliser le Dublin Core Qualifié et propose un tableau reprenant les éléments DC et expliquant pour chacun comment préparer ses métadonnées et quels vocabulaires et listes contrôlés appliquer. En outre, le guide propose un *mapping* ou tableau d'équivalence entre les formats MARC et UNIMARC et le format DC Qualifié. TEL a mis en place ses propres listes contrôlées pour certains éléments (l'élément *droit* par exemple). Ces listes sont vivement recommandées mais non obligatoires.

- **OAIster**

  OAIster est le « vétéran » des portails sélectionnés. Son interface permet de rechercher parmi plus de 13 millions de notices et d'accéder (selon les droits afférents) aux ressources numériques correspondantes.

  Un lien vers le site de la DLF permet d'accéder à un wiki de bonnes pratiques concernant l'OAI. La section qui nous intéresse plus particulièrement concerne le travail sur les métadonnées. En ce qui concerne les formats, outre le Dublin Core, OAIster conseille aux fournisseurs de données de proposer leurs métadonnées sous d'autres formats de description également.

4 Résultats de l'étude

Cette partie est une synthèse des observations effectuées sur différents aspects de la recherche communs aux sites étudiés. Elle est volontairement sélective et a pour but de donner un aperçu de la façon dont quatre portails différents gèrent les fonctionnalités de recherche à partir de sources hétérogènes. Même si des questions d'usabilité* sont abordées, le positionnement en tant que documentaliste permet, à partir de l'étude des points d'accès et des métadonnées, de déduire un certain nombre de caractéristiques du fonctionnement en *back-office*.

Les résultats de l'étude réalisée sur les fonctionnalités de recherche ont été classés en trois catégories principales auxquelles feront écho des préconisations en matière d'interopérabilité sémantique, syntaxique et organisationnelle en troisième partie. La première sous-partie s'attache à observer le traitement de la **granularité**, la deuxième, à évaluer le **degré de cohérence** à l'œuvre dans le portail alors que la dernière traite plutôt de la question d'**usabilité*** en s'attachant au chemin que l'utilisateur parcourt pour accéder à la ressource numérique. Ces trois focales ont été choisies car elles permettent de mettre l'accent sur des points de friction qui pourraient exister dans le développement de portails intégrés tels que sous OAI-PMH.

Construire un portail, c'est créer à partir de données et de sources hétérogènes un accès simplifié présentant une certaine homogénéité. Il est en effet essentiel pour le confort de l'utilisateur de lui fournir des repères stables lui permettant de comprendre d'instinct comment l'interface est organisée ; les points concernant la granularité et le chemin d'accès au document numérique sont à ce titre révélateurs. L'évaluation de la cohérence et de la transversalité par l'étude des métadonnées est, elle, au cœur des problématiques de la mise en place d'un portail en OAI-PMH : le protocole définissant des standards pour l'échange des métadonnées, évaluer comment ces métadonnées sont organisées au sein d'un portail, quelles passerelles il existe entre elles et quels points d'accès permettent d'accéder aux ressources revient à mettre en lumière les aspects que le protocole OAI-PMH ne règle pas.

4.1 Traitement de la granularité : l'étude au niveau de la notice

4.1.1 Introduction

Tout d'abord, voici une définition du concept de granularité de l'information : c'est le **niveau de description d'une unité d'information**.

> ➔ Prenons l'exemple d'un coffret de CD regroupant l'œuvre d'un compositeur. On peut choisir comme unité d'information à décrire le coffret, un disque, une œuvre au sein d'un disque, un mouvement d'une œuvre… On peut également décrire plusieurs niveaux et les lier entre eux : cela amène à créer des notices mères et des notices filles, qui s'emboîtent comme des poupées russes.

Selon les objectifs que l'on a mais aussi selon les utilisateurs, on choisira un niveau de granularité plus ou moins élevé (*cf Préconisations p. 75 pour plus de précisions*).

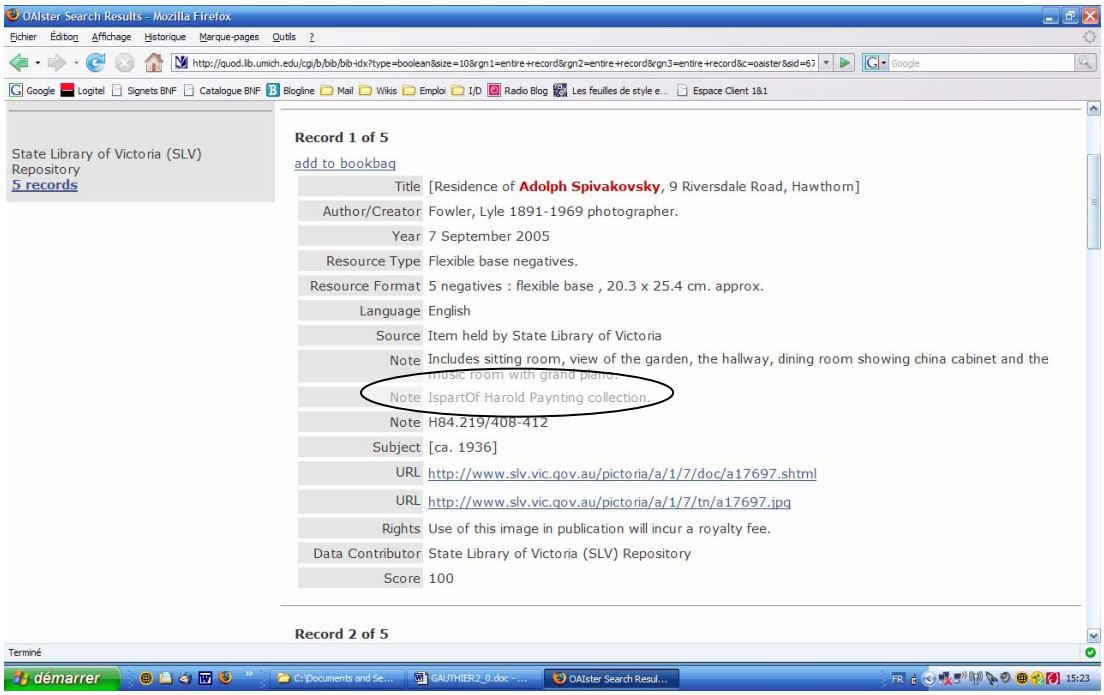
Sur un portail par nature intégré, il est important pour la bonne compréhension et pour l'exploitation de l'information par l'utilisateur de pouvoir replacer une ressource dans son **contexte**. Or, dans le cas d'une granularité très fine, l'utilisateur doit pouvoir être en mesure de lier une « partie » à son « tout ».

4.1.2 Résultats

Il n'existe apparemment pas de notices mères-filles dans les portails étudiés, seulement une **mention de relation**, sans lien à une autre notice, permettant de **contextualiser l'information**.

OAIster utilise un champ *Note* pour indiquer l'appartenance d'une ressource à un ensemble plus grand. La valeur du champ, selon le fournisseur de données, est plus ou moins claire : elle va de la simple URL à la description d'une collection. Quant à la notice décrite, on ne comprend pas toujours clairement à quelle niveau de granularité on se trouve.

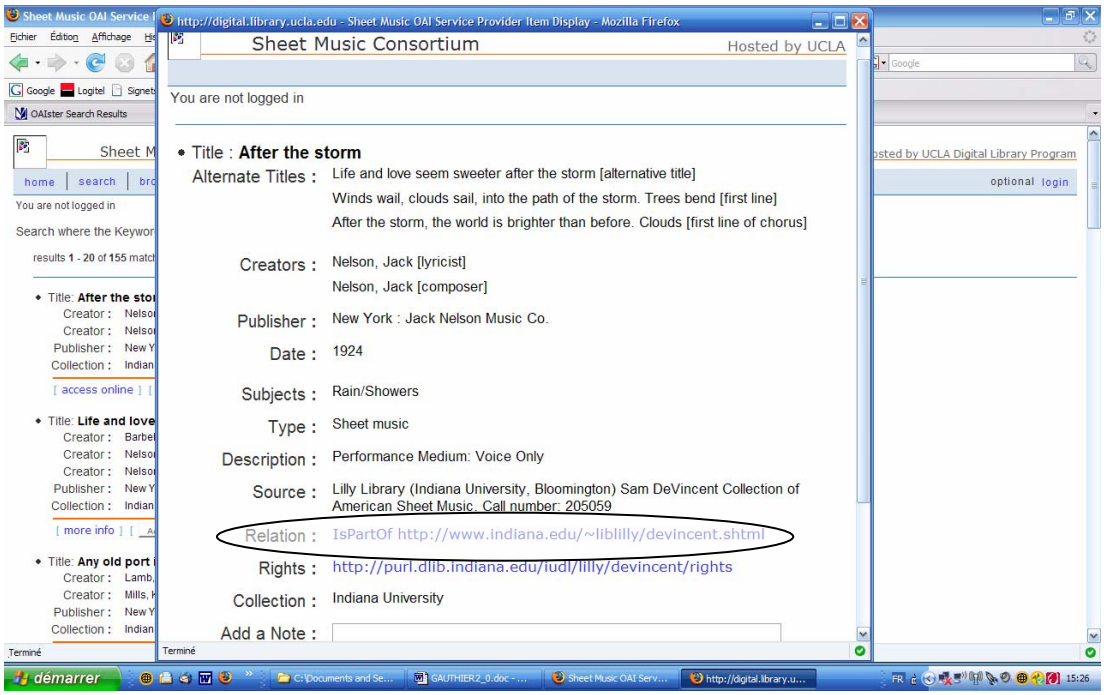
> ➔ Ainsi, certaines notices décrivent des bases de données ou des sites web, ce que l'on découvre en sortant d'OAIster lorsqu'on clique sur le lien menant à la ressource.
>
> Pour la plupart, la description se fait au plus fin (par exemple photographie avec mention très succincte en *Note* de la collection à laquelle elle appartient pour les notices de la State Library of Victoria). Ce niveau de granularité simple, sans lien entre notices mères et filles, a dû être privilégié par OAIster pour répondre aux attentes d'un public assez large et diversifié.



Notice OAIster d'une photo numérisée provenant de la State Library of Australia avec mention de la collection d'appartenance en note

Concernant TEL, il n'existe apparemment pas de champ indiquant à quel niveau de granularité on se trouve. L'indication du contexte nécessaire dans la compréhension de la ressource n'est souvent pas présente.

Sheet Music Consortium et Europeana ne décrivent les ressources qu'à un seul niveau de granularité : la partition et le livre numérisé. Pour Sheet Music Consortium, on trouve pour certains FD mention de relation avec un lien URL à la collection à laquelle appartient la ressource.



Notice Sheet Music Consortium avec lien vers la collection d'appartenance

4.1.3 Ce que l'on peut retenir

Le manque de **régularité** dans l'indication du niveau de description de l'unité d'information dans les deux portails généralistes peut nuire à la bonne compréhension et par là même à l'exploitation de l'information trouvée par l'utilisateur.

Les liens entre des notices décrivant différents niveaux d'une même ressource semblent être intéressants à conserver dans un accès intégré vu qu'ils permettent de **contextualiser** l'information.

En règle générale, la présence d'un simple lien pour matérialiser une relation semble nécessaire mais pas suffisante. Peut-être serait-ce intéressant, à l'instar de certains FD d'OAIster, de fournir une note concise sur la collection d'appartenance ?

4.2 Evaluation de la transversalité : étude au niveau du champ

4.2.1 Introduction

Pour Shreeves (20, Shreeves), une certaine homogénéité dans les métadonnées permet de rendre ces dernières « **partageables** » et donc **exploitables** dans un environnement intégré tel qu'un portail.

Elle répertorie les résultats de diverses études menées sur la question de la qualité et de la « partageabilité » des métadonnées. Elle note qu'il existe souvent un :

- **Manque de cohérence** à l'intérieur d'un même lot (utilisation des éléments *date* et *couverture* dans un même enregistrement par exemple) ;
- **Trop-plein d'informations** (ajout d'informations inutiles telles que le type de scanner utilisé pour numériser une image) ;
- **Manque d'informations essentielles** à la compréhension des métadonnées (ne pas fournir le nom de la collection à laquelle une image appartient, privant ainsi l'utilisateur d'une donnée contextuelle essentielle) ;
- **Non-respect des standards** techniques.

Des métadonnées de qualité sont des métadonnées que l'on peut utiliser dans un environnement intégré. Elles doivent donc s'adapter au contexte d'utilisation (concept de modularité).

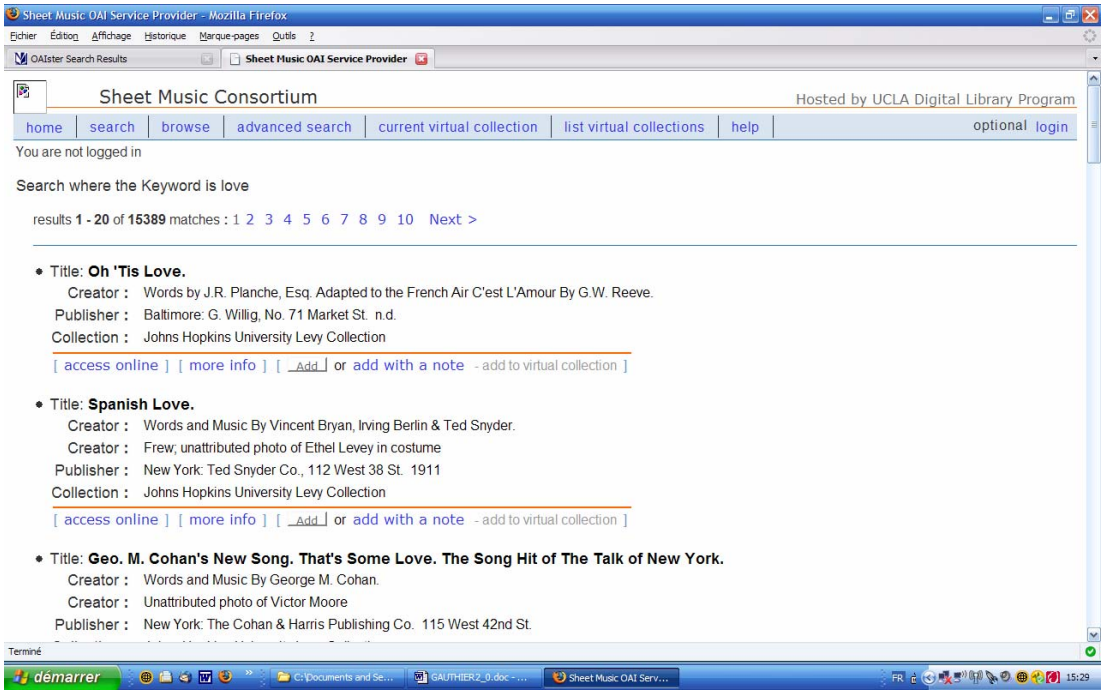
L'étude de l'affichage des résultats permet d'obtenir un aperçu de la qualité des métadonnées proposées par les FD et traitées par les FS.

4.2.2 Les champs

La plus forte hétérogénéité au niveau des champs affichés se retrouve dans les portails généralistes encyclopédiques. Ces portails ont à gérer un nombre important de FD et donc de notices et de métadonnées, ce qui rend la tâche d'harmonisation complexe. Le portail Sheet Music Consortium réussit à présenter assez clairement ses notices provenant d'horizons divers. Au niveau de la présentation et de la clarté des notices, Europeana se trouve entre les deux ; le résultat n'est pas toujours homogène mais l'utilisateur s'y retrouve quand même. Il convient tout de même de noter qu'Europeana n'agrège les métadonnées que de trois FD différents, ce qui rend la tâche d'harmonisation et la lecture des résultats plus aisées.

4.2.2.1 L'affichage liste

➔ **Sheet Music Consortium** propose un affichage des éléments plutôt homogène (*titre*, *créateur*, *éditeur*, *collection* toujours présents et dans le même ordre), avec une indication claire de la provenance de la notice, ce qui paraît essentiel. En revanche, il peut y avoir un seul champ *creator* comme il peut y en avoir 8 (par exemple *compositeur*, *parolier*, *artiste*, *groupe*, etc...), ce qui peut nuire à la lecture des notices en affichage liste.



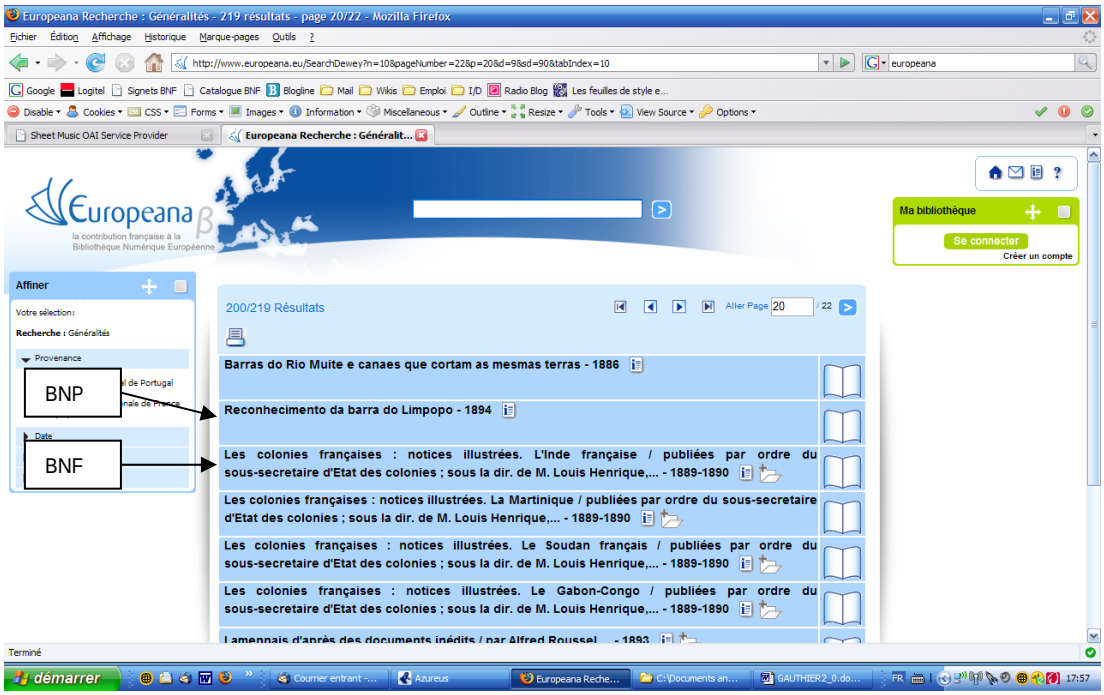
Affichage liste sur Sheet Music Consortium

➔ **Europeana** : l'affichage du format liste n'est pas toujours le même (*titre*/*auteur* parfois/*date*), les champs *titre* et *auteur* ne sont pas toujours séparés par le même séparateur. Seule constante : la date est toujours affichée au même format et séparée des autres données par un tiret.

Ex. : Sonetos completos – 1934

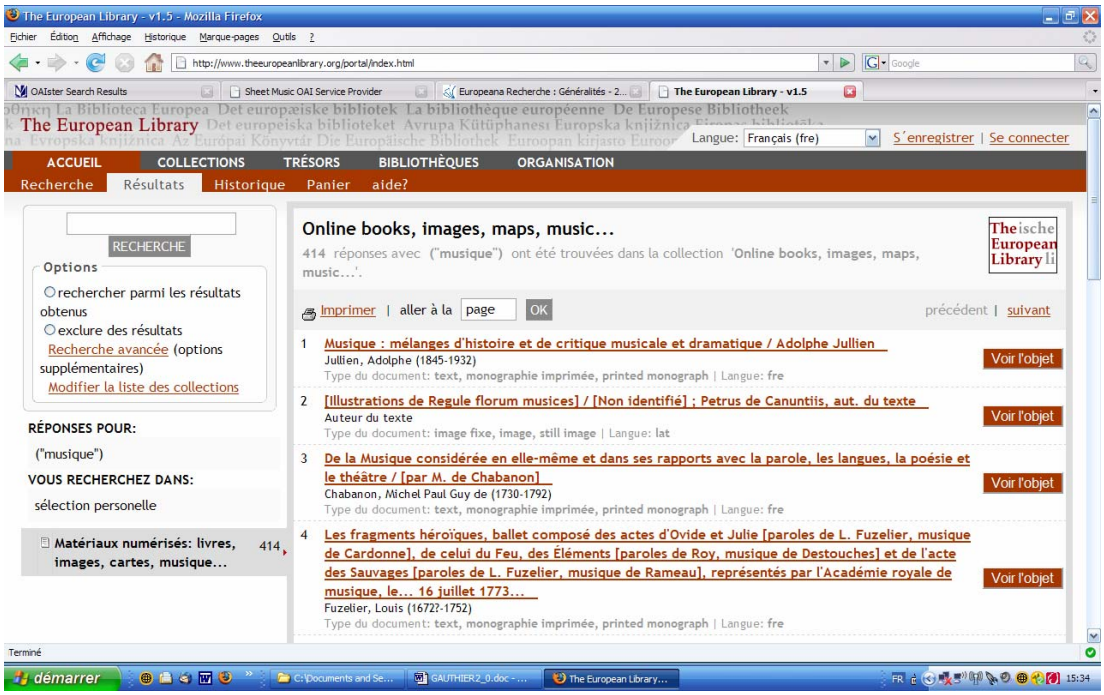
Symphonies et poèmes / V. de Laprade - 1920

En affichant la notice complète, on s'aperçoit que la notice allégée est le résultat d'une concaténation des champs *titre* et *date* et que la BNF inclut l'auteur dans son champ titre (séparés par un slash), ce qui explique l'apparition de l'auteur dans l'affichage en liste. On peut se demander si l'apparition de l'auteur à la fois dans le titre et dans le champ *auteur* n'est pas préjudiciable pour l'exploitation des métadonnées par le FS et pour la recherche...



Affichage liste sur Europeana avec notices de la BNF et de la Bibliothèque du Portugal

➔ **TEL** : l'affichage en liste est assez homogène et lisible (*titre* sous forme de lien actif menant à la notice complète, *auteur*, *type* de ressources et langue).



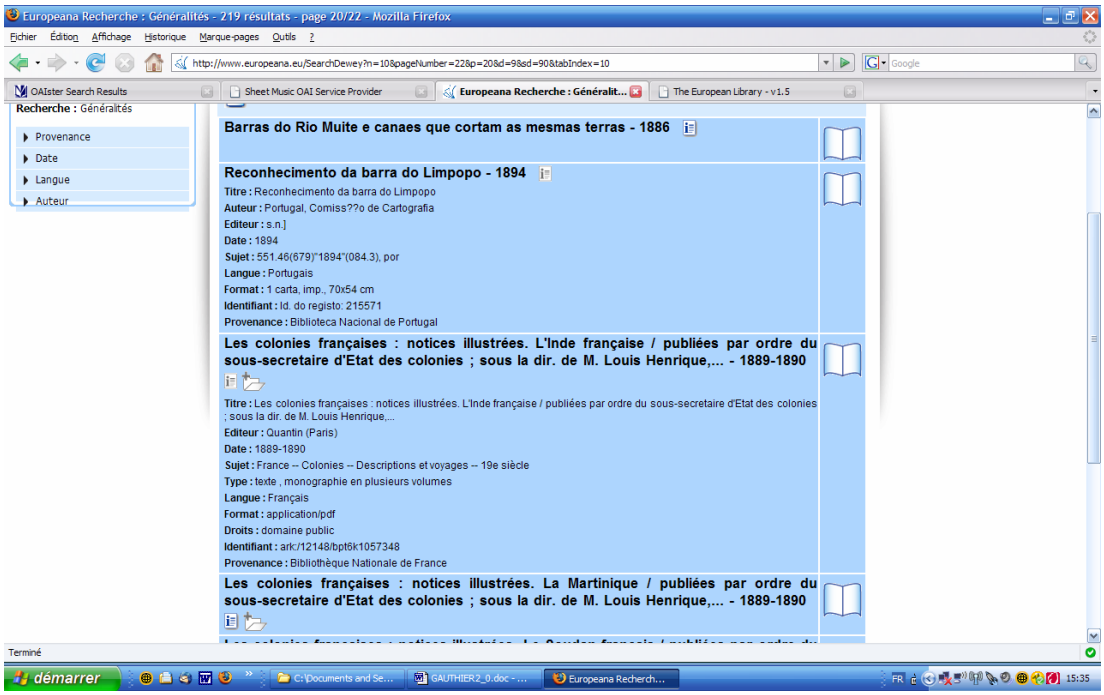
Affichage liste sur TEL

➔ **OAIster** ne présente pas d'affichage liste, ce qui pose problème au niveau de l'appréhension des résultats par l'utilisateur, qui ne peut avoir une vue d'ensemble du lot issu de la recherche effectuée.

4.2.2.2 L'affichage complet

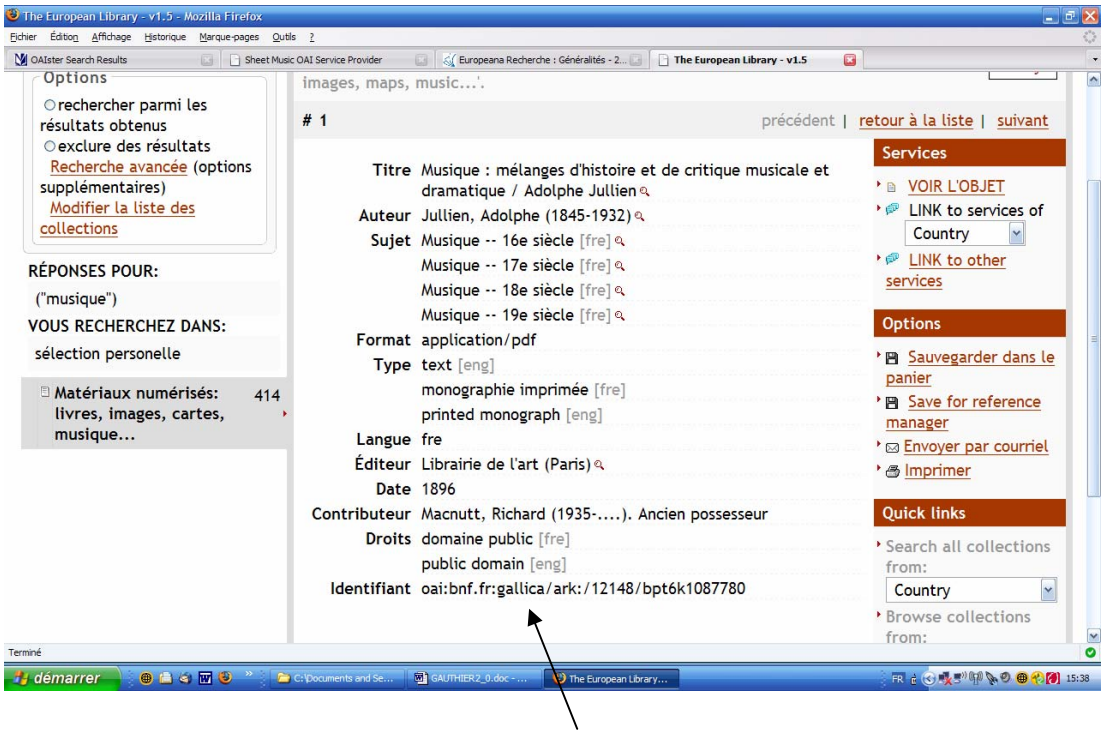
Le nombre de champs utilisés pour décrire les ressources varie dans chacun des portails selon le FD.

➔ Ainsi, sur **Europeana**, les notices BNF font toujours mention des droits (mention *domaine public* sur toutes les notices), contrairement à celles des deux autres bibliothèques.

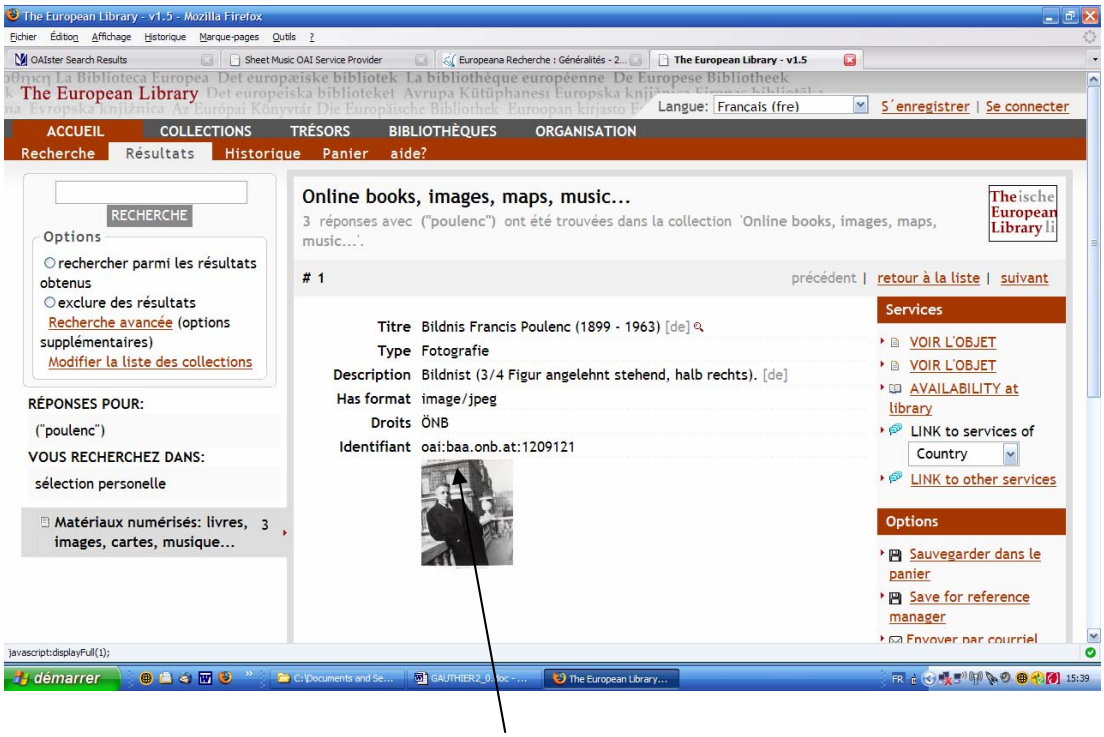


Affichage de notices de la Bibliothèque du Portugal et de la BNF

➔ Chez **TEL**, l'affichage de la notice complète n'est pas homogène : il varie très grandement selon le FD. Un gros problème : on ne sait pas toujours, lorsqu'on recherche dans *Matériaux numérisés*, à quel FD appartiennent les ressources… L'identifiant nous fournit l'information mais de façon peu claire.



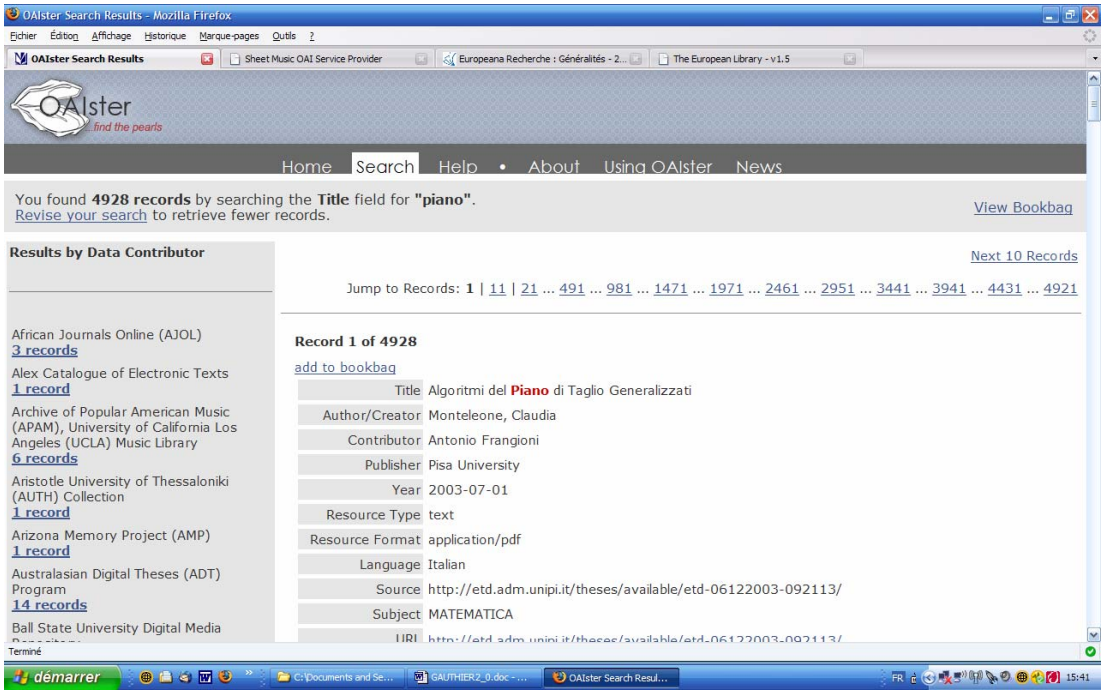
Affichage sur TEL d'une notice provenant de Gallica, la bibliothèque numérique de la BNF



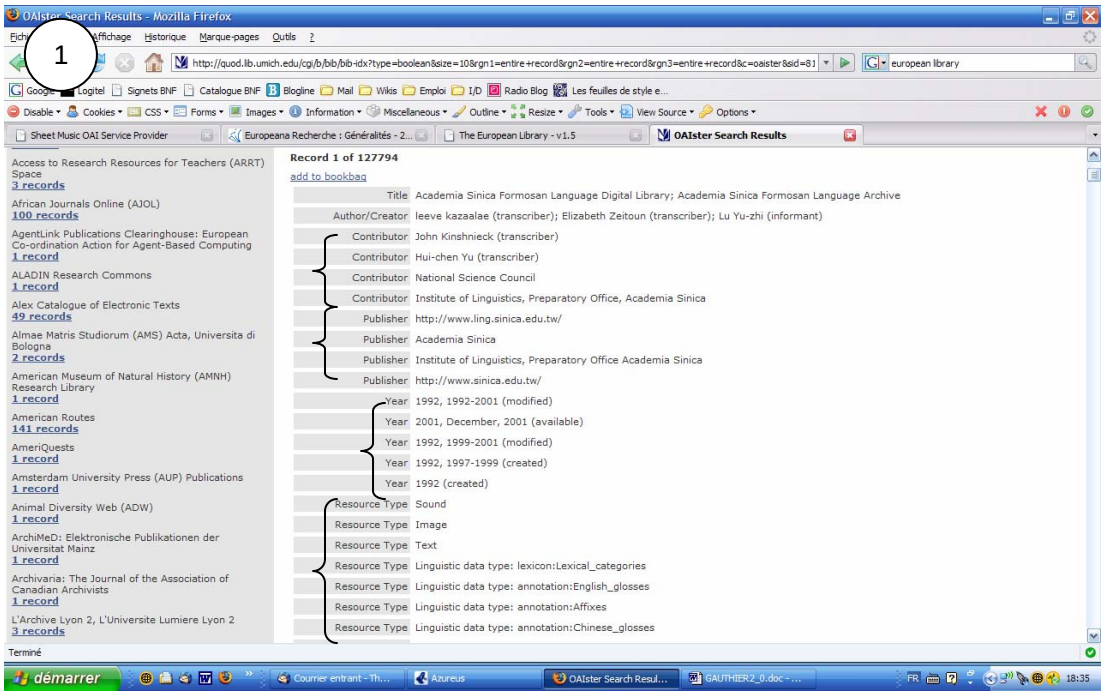
Affichage sur TEL d'une notice de la Bildarchiv Austria

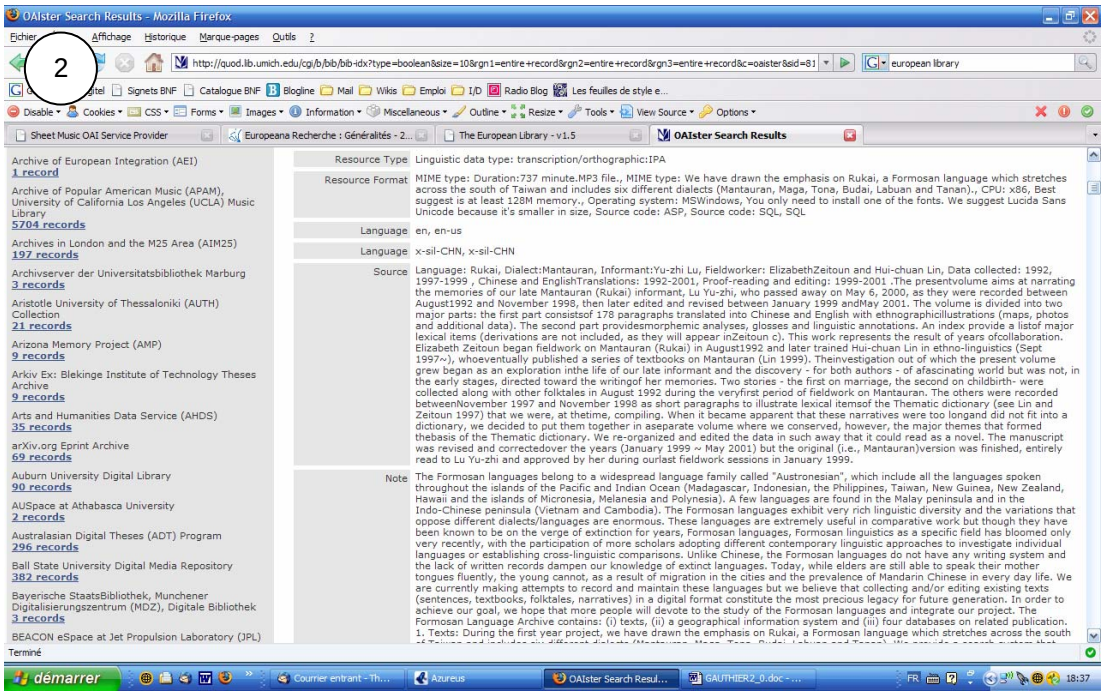
➔ **OAIster/DLF** : seul l'affichage complet existe, ce qui rend l'appréhension des résultats difficile pour l'utilisateur. L'hétérogénéité y est très forte : certaines notices présentent très

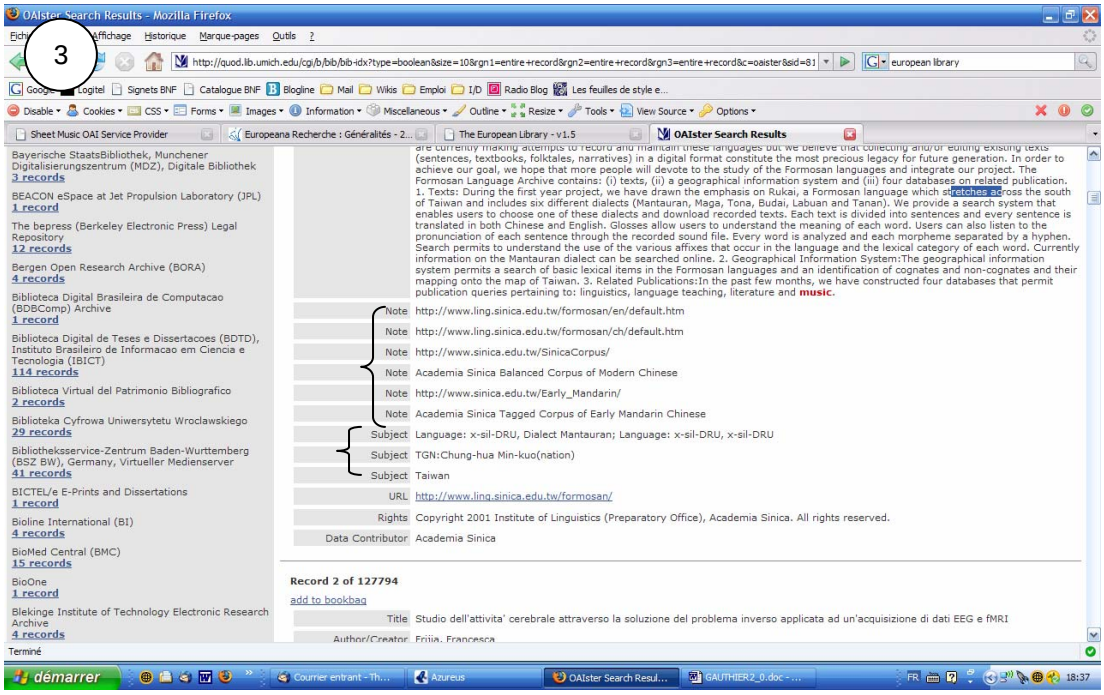
peu de champs, d'autres en ont un nombre imposant (ce qui pose problème à la lecture), avec parfois le même élément répété plusieurs fois (notamment l'élément *Note*, qui peut être répété un grand nombre de fois). Le bon point : on sait d'où proviennent les données et on peut choisir à tout moment de ne consulter que les ressources d'un FD.



Affichage sur OAIster de la page de résultats de la recherche « piano » sur les mots du titre







Affichage d'une notice en trois parties (recherche avec le terme « music ») avec champs répétés de nombreuses fois (notamment l'élément Note)

Il est à noter que sur les quatre portails, l'affichage de la notice complète se fait sous la forme d'une liste de champs (*titre*, *auteur*, etc…) avec en regard leurs valeurs respectives. En outre, certains types de champs (*auteur*, *note*, *sujet*) sont parfois répétés plusieurs fois.

Cette présentation assez simplifiée peut devenir gênante dans l'appréhension des résultats par l'utilisateur, surtout lorsque la notice possède un volume d'information important. Une présentation par concaténation de certains champs permettrait un affichage plus clair.

4.2.3 Les valeurs des champs

Il est entendu par valeur du champ ou de l'élément (qui est l'appellation propre au Dublin Core) l'**information qui lui est assignée**. Ainsi le champ *auteur* peut avoir pour valeur *Hugo, Victor*.

4.2.3.1 Règles d'écriture : les champs date, titre et auteur

Pour que l'indexation des champs et par là même pour que l'interrogation soient efficaces, il est préférable que les champs d'un même type reçoivent le même type d'information et suivent les mêmes **règles d'écriture**. Cet aspect déjà difficile à mettre en place en local, apparaît très problématique lorsqu'on réalise un portail en OAI-PMH. Là encore, les portails les plus importants en terme de volume d'information traité, présentent le plus de problèmes au niveau syntaxique de leurs métadonnées et ce, même si un guide de bonnes pratiques existe. Une étude réalisée sur le sujet a catégorisé les problèmes rencontrés sur les métadonnées (17, Hillmann). Voici deux catégories qui illustrent cette partie :

- **Données inexactes** : la valeur d'un élément ne correspond pas au format ;
- **Données imprécises** : plusieurs valeurs sont entrées là où une seule valeur doit être.

L'étude des champs *date*, *titre* et *auteur* vont permettre de déceler les problèmes syntaxiques les plus couramment rencontrés.

Le site **Sheet Music Consortium** présente le plus de cohérence à ce niveau-là. Par exemple, le champ *creator* est à peu près normalisé (Nom, Prénom, avec parfois les dates biographiques et un qualificatif) et les ambiguïtés concernant la valeur de certains champs (comme le champ *creator*) qui pourraient exister sont souvent levées grâce à l'existence d'un guide à propos du format DC et de l'utilisation de qualificatifs (*creator* peut ainsi recevoir comme qualificatif *lyricist* ou *composer* par exemple). Il est à noter également que c'est le seul portail à proposer une recherche précise et efficace par date.

Pour le reste, l'aspect syntaxique pose de nombreux problèmes qui seront illustrés à l'appui des champs *date*, *titre* et *auteur* selon le portail.

➔ L'élément *date* chez TEL se trouve sous des formes très variées, rendant difficile le développement d'une recherche future à partir de la date :

195-

1866-1868

18.

1926

1917]

Non seulement il n'y a pas normalisation au niveau de l'écriture du champ mais il reste également des traces des anciens formats utilisés (le crochet après *1917*).

➔ L'élément *titre* chez Europeana peut poser également problème :

Souvenirs d'un demi-siècle. 1, Au temps de Louis-Philippe et de Napoléon III, 1830-1870 / Maxime Du Camp

La nouvelle Carthage : roman / Georges Eekhoud

Les bases de la morale évolutionniste / par Herbert Spencer

La mention d'un auteur ou d'un type de document dans le titre est ce que Tennant nomme en anglais *artifact* (33, Tennant). Ce sont des « traces » du format original (par exemple MARC) laissées lors de la migration des métadonnées vers un autre format. On peut ranger l'*artifact* dans la catégorie « données imprécises » mentionnée plus haut.

Dans notre exemple, l'élément *titre* sur Europeana correspond à l'élément 200 *Titre et mention de responsabilité* du format MARC ; il y a eu concaténation des champs 200$a (*titre*) et 200$f (*auteur principal*).

➔ On trouve les mêmes difficultés de traitement sur l'élément *auteur/ créateur* sur OAIster :

Niles, John Jacob, 1892-1980

Unknown;

Unknown

Hudson Photo

Henning, Paul

leeve kazaalae (transcriber); Elizabeth Zeitoun (transcriber); Lu Yu-zhi (informant)

Ces imprécisions, ce manque de cohérence non seulement peuvent **gêner à la lecture** des résultats mais aussi rendre la **recherche moins précise**, surtout en cas de portails moissonnant un grand nombre de métadonnées. **Ainsi, pour réduire le bruit à la recherche, il est nécessaire de réduire les imprécisions**.

4.2.3.2 Points d'accès

Les points d'accès* sont les **critères de recherche** (*auteur*, *sujet*, …) qui permettent à l'utilisateur d'accéder à une ressource. *Cf Formulaires de recherche en annexe p. 130 à 135.*

Pour assurer la transversalité entre les notices de divers FD, il faut, en plus d'une certaine homogénéité au niveau des règles d'écriture, des « **passerelles** », des vocabulaires communs pour lier les notices entre elles. Ces passerelles, qui permettent toute une richesse et une précision au niveau de l'interrogation de certaines bases de données du domaine patrimonial, proviennent en particulier d'un travail au niveau des notices d'autorité*. La notice d'autorité est une donnée structurée (elle peut concerner une personne, un titre, une œuvre, un descripteur) indépendante des notices bibliographiques qu'elle décrit ; elle est maintenue dans une base de données* ou une table/application différente. Cela présente l'intérêt de **normaliser** ces fichiers - en levant par exemple les ambiguïtés liées à la synonymie du langage naturel ou aux variations orthographiques d'un nom d'auteur – et de rendre la **recherche plus puissante** car plus précise.

L'étude des simples fonctionnalités de recherche du côté de l'interface ne permet pas de dire s'il y a utilisation de vocabulaires communs ou tout du moins, mise en place d'équivalence entre les vocabulaires contrôlés utilisés par les divers acteurs de chaque portail. En

revanche, elle permet de voir quels **points d'accès** mènent à la transversalité inter-collection.

- **Type de ressources**

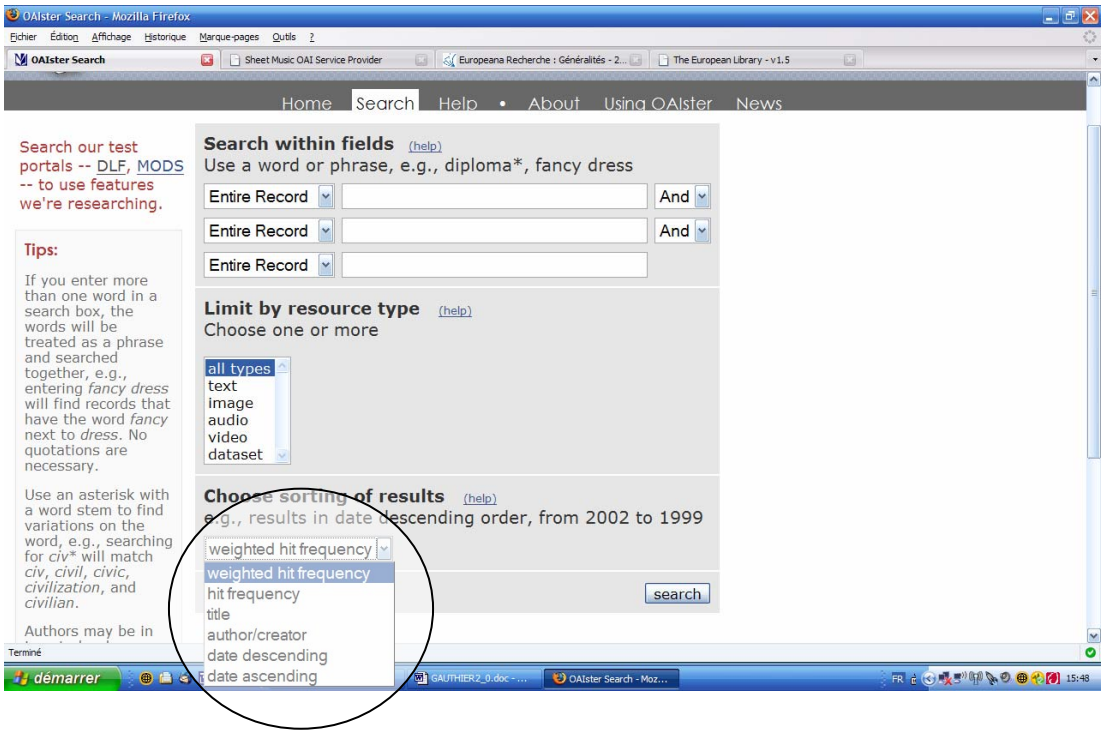
  Le type de ressources est l'élément le plus normalisé : les quatre portails permettent de rechercher par type de ressources parmi toutes les notices moissonnées. Cela est dû au fait qu'il y a un nombre très limité de types, facilitant ainsi le traitement de ce champ par le fournisseur de données et le fournisseur de services.

- **Langue**

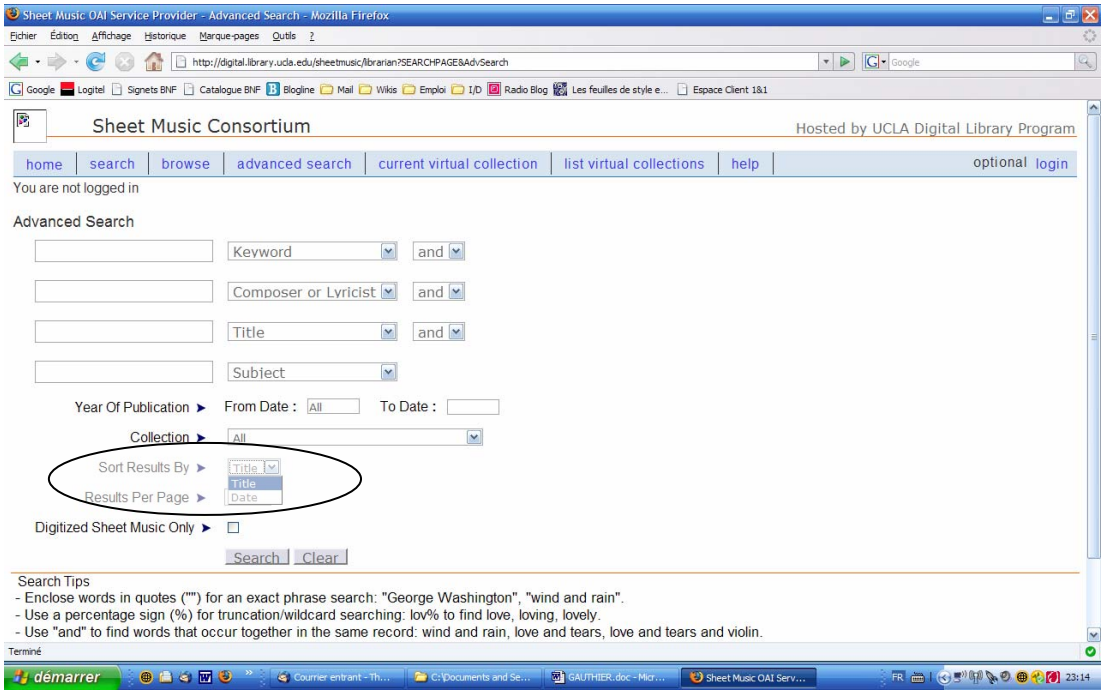
  Europeana et TEL ont tous deux normalisé le champ *langue*. La recherche à partir de la langue du document est efficace sur les deux portails. Ce critère semble important dans un environnement intégré ; en effet, l'apparition de notices dans une langue qu'on ne lit pas est gênante pour l'exploitation des résultats.

- **Possibilités de tri**

  Deux portails proposent dès le formulaire de recherche des options de **tri** des résultats : OAIster et Sheet Music Consortium. Sur le premier, on peut choisir de classer les résultats selon l'un des critères suivants : l'auteur/contributeur, la date, le titre, la pertinence. Sur Sheet Music Consortium, on peut classer selon la date et le titre.



Options de tri dès la recherche sur OAIster

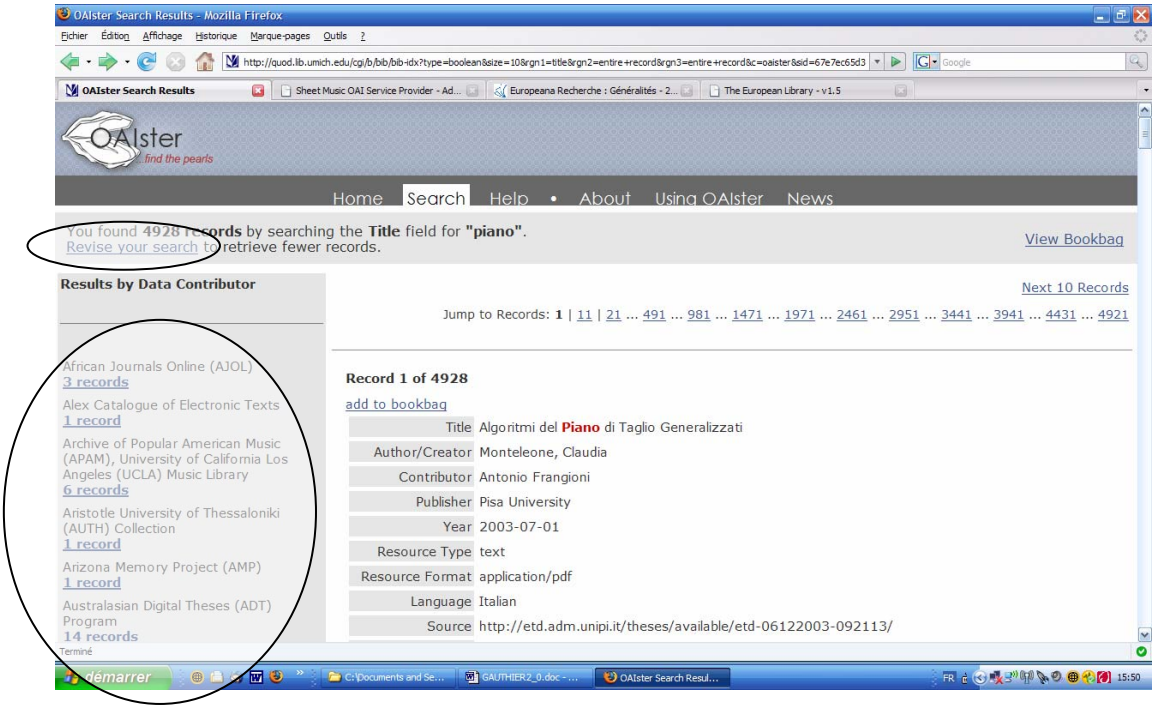


Options de tri dès la recherche sur Sheet Music Consortium

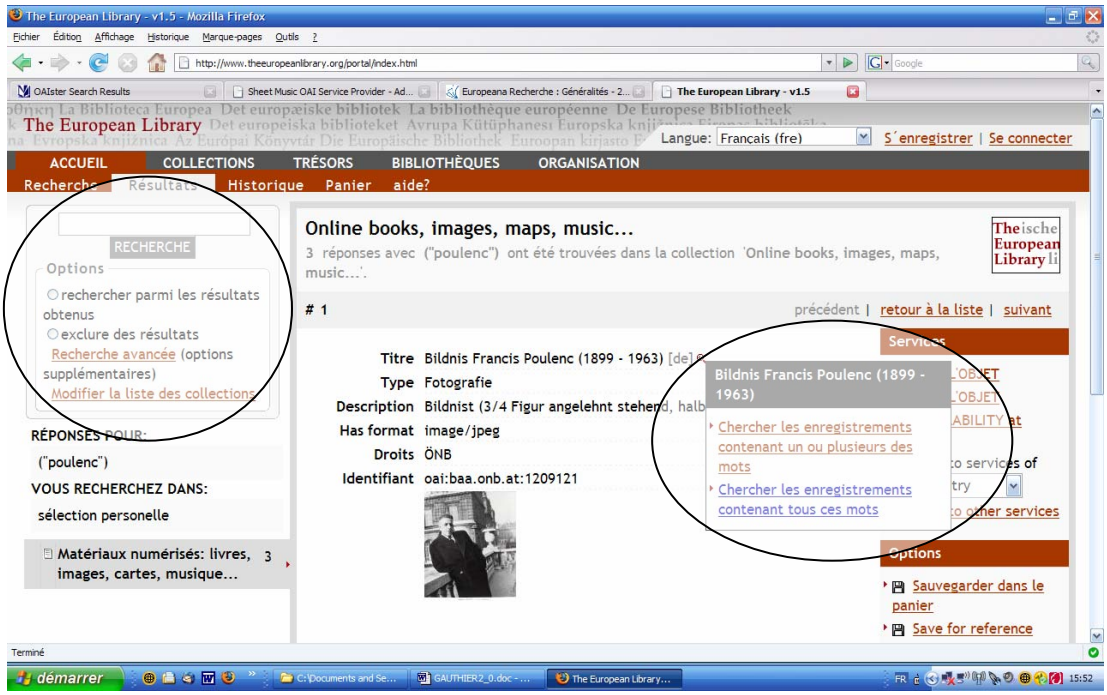
Les résultats sur OAIster ne sont pas très concluants ; le classement par auteur n'a pas l'air de fonctionner ; le champ date n'étant pas normalisé pour tous les FD, le classement est

approximatif. Sheet Music propose quant à lui des tris assez pertinents selon les deux critères.

Tous les portails à l'exception de Sheet Music Consortium proposent d'**affiner** les résultats. Sur OAIster, les résultats peuvent être affichés selon le FD, ce qui est très pratique. Sur TEL, il est possible d'effectuer à nouveau une recherche sur les mots du titre, de l'auteur et de l'éditeur. C'est une autre façon d'accéder aux fonctions de recherche avancée.



Affichage des résultats selon le FD et lien Revise your search *ramenant au formulaire de recherche sur OAIster*



Options d'affinage sur TEL

Quant à Europeana, ses possibilités de tri sont examinées dans la partie qui suit.

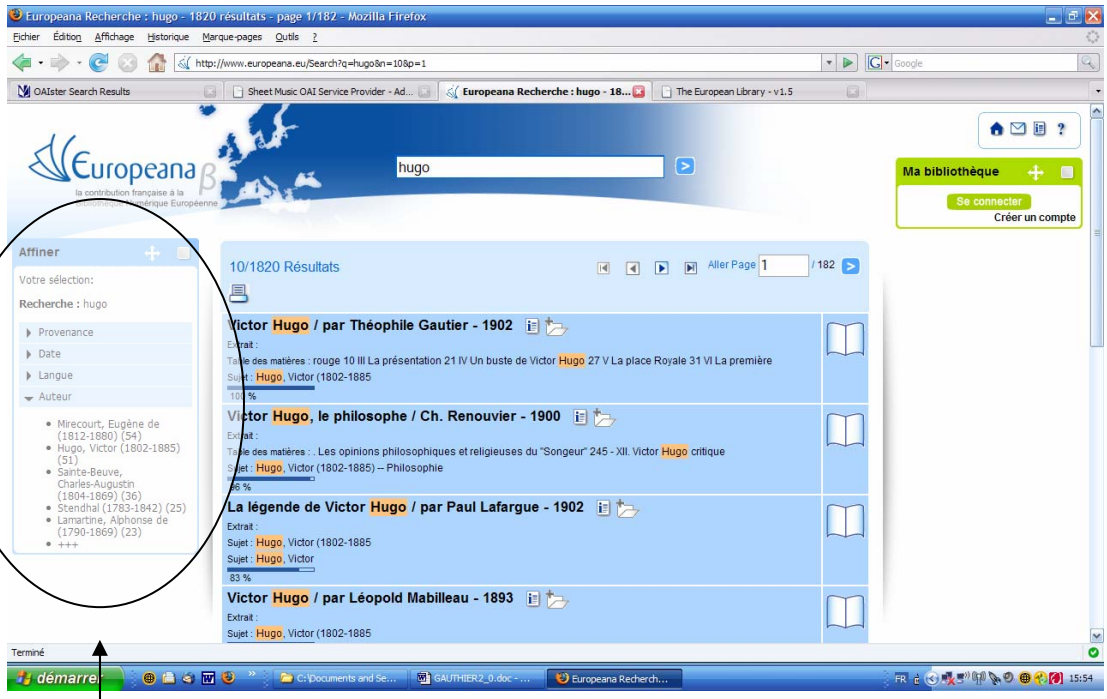
4.2.3.2.1 Europeana, une transversalité réussie

Parmi les portails étudiés, Europeana offre les meilleurs résultats en termes de **transversalité**. Tout d'abord, les notices ont été classées selon un thème général (Généralités, Philosophie/Psychologie, Religion,...), chaque thème étant lui-même divisé en sujets (Philosophie antique, philosophie moderne pour le thème Philosophie/Psychologie par exemple ➔ *cf Formulaire de recherche p. 131*). Cette pré-sélection par sujet permet une transversalité entre les trois bibliothèques partenaires. En outre, il est également possible de rechercher selon le siècle de parution (du 16ème au 20ème s.), la langue de la ressource et la provenance de la ressource. A l'affichage des résultats, une palette sur la gauche permet de combiner ces critères pour affiner la recherche. Ainsi, on peut commencer une recherche selon un thème puis affiner les résultats selon la langue.

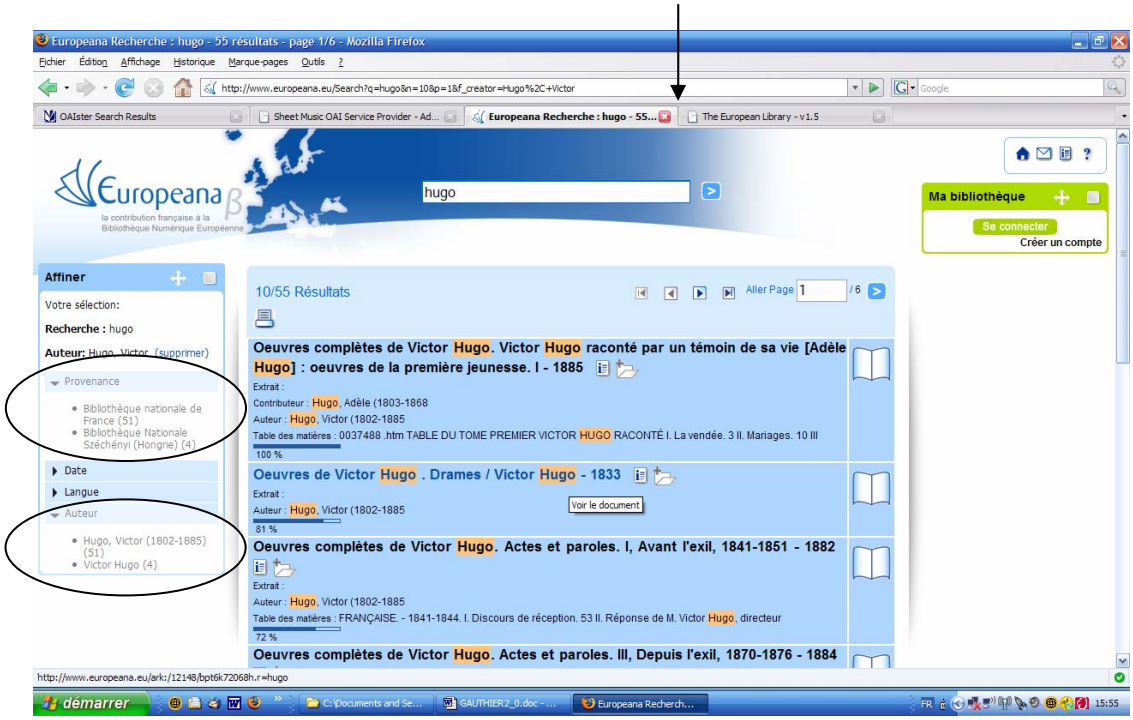
Le travail sur l'élément *auteur* est également intéressant. Il n'est pas possible de rechercher directement par un auteur (les points d'accès sur Europeana sont une recherche plein texte sur mots des champs, une entrée par critères et une entrée par thèmes). En revanche, à l'affichage des résultats, l'onglet présent sur la gauche de l'écran permet de faire dérouler un index d'auteurs (très peu pratique au niveau de la navigation) interrogeables. La recherche suivante a été effectuée :

➔ « hugo » dans le champ de recherche simple

➔ En cliquant sur « Hugo, Victor (1802-1885) » dans l'onglet de gauche, on obtient 51 réponses avec cet auteur dans la BNF et 4 réponses sous l'auteur « Victor Hugo » dans le fonds de la Bibliothèque Nationale Széchényi. Il y a donc bien traitement des autorités dans ce portail.



Palette d'affinage avec index d'auteurs sur Europeana / Deux écritures de l'auteur Victor Hugo selon le FD



Même si la navigation n'est pas toujours aisée, il y a possibilité d'effectuer des recherches transversales sur un auteur dans Europeana. Il convient tout de même de noter qu'Europeana, en tant que prototype au nombre réduit de notices, n'offre pas beaucoup de résultats. Son étude est surtout intéressante au niveau des possibilités qu'il offre en terme de recherche. Son accès par grands thèmes l'apparente au niveau de la recherche aux répertoires que l'on peut trouver sur le web.

4.2.3.2.2 Recherche sur les mots du champ

Pour les autres portails, la recherche se fait apparemment sur les **mots des champs** : le moteur indexe le contenu d'un ou plusieurs des champs des notices moissonnées et la recherche s'effectue sur les mots de cet index.

Sheet Music Consortium propose une recherche sur les points d'accès *Composer/Lyricist* (qualificatifs du champ DC Creator), *Title*, *Subject* et *Publisher*. La recherche par thèmes s'effectue sur les mots du champ *Subject* indexé par le moteur (*cf Formulaire en annexe p. 134*).

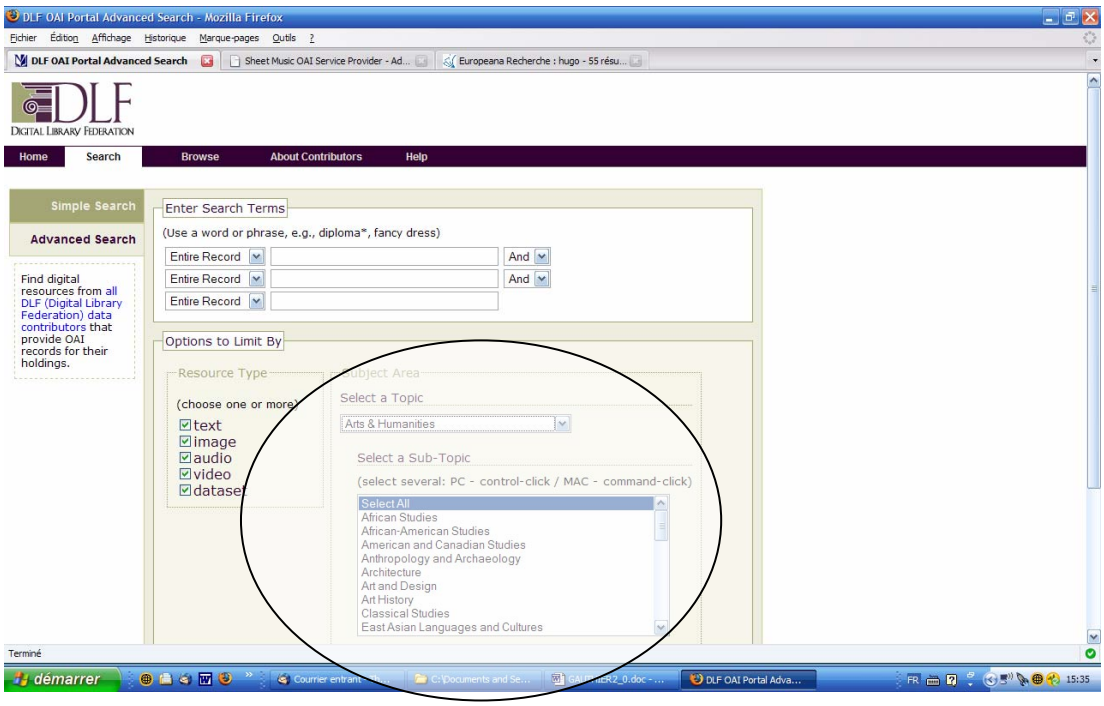
La recherche avancée chez TEL et OAIster (*cf formulaires pp. 132 et 135*) peut se faire sur différents champs combinés entre eux par des opérateurs booléens mais elle reste une recherche en texte intégral, c'est-à-dire sur un vocabulaire qui n'est pas normalisé. Pour un site déjà homogène dans son contenu et sa forme comme Sheet Music Consortium, cela ne nuit pas réellement à la recherche. En revanche, sur les deux sites les plus importants, la recherche produit beaucoup de bruit.

4.2.3.2.3 Les particularités de la DLF

Les personnes responsables du développement d'OAIster se sont aperçues que plus le nombre de fournisseurs de données et donc de ressources grandissait, plus la recherche devenait inefficace. Ces faiblesses au niveau de la recherche les ont donc conduit à développer un prototype, le site DLF, afin de tester de nouveaux outils. La DLF développe ainsi le **cluster**, qui permet de **catégoriser un très grand nombre de données** à la fois. En effet, s'il est possible de créer ou d'utiliser des vocabulaires contrôlés sur un fonds limité en nombre et dans les domaines traités, la tâche se révèle pratiquement impossible sur un très grand nombre de notices ou sur le web.

C'est ainsi que le site de la DLF permet une entrée par thèmes (Arts and Humanities, Science, Engineering, etc…) et sous-thèmes. Mais cette technique n'est pas encore rôdée : le classement n'est pas toujours concluant (une photographie d'une vieille femme à la guitare de Dorothea Lange est indexée entre autres par les termes *Electrical Engineering* et

Infertility !). Selon Hagedorn (28, Hagedorn), la catégorisation en sciences humaines est beaucoup plus imprécise qu'en sciences exactes car le langage dans ce domaine présente plus d'ambiguïtés (par l'utilisation de métaphores par exemple), rendant la catégorisation plus complexe et délicate à effectuer.



Formulaire de recherche avancée sur la DLF avec pour points d'accès remarquables le domaine *et le* sous-domaine

4.2.4 Ce que l'on peut retenir

On voit qu'il se dessine une **ligne de partage** entre des portails bien balisés, dont les ressources appartiennent au même type (Sheet Music Consortium) et/ou proviennent du même domaine professionnel (Europeana) et les portails à vocation encyclopédique. Dans les premiers, la recherche plein texte sur mots d'un ou plusieurs champs (Sheet Music Consortium) est efficace et intéressante car une certaine cohérence entre les fournisseurs de données existe a priori. Cette même recherche plein texte dans des portails à très grande échelle produit énormément de bruit, ce qui a conduit OAIster à tenter une catégorisation a posteriori de ses notices. Cette classification, qui permet un accès de type répertoire, n'offre pas les mêmes résultats selon le type de portail. Europeana parvient à réaliser une classification plutôt pertinente des notices moissonnées alors qu'OAIster/DLF rencontre encore de nombreuses difficultés.

4.3 Le chemin d'accès au document numérique

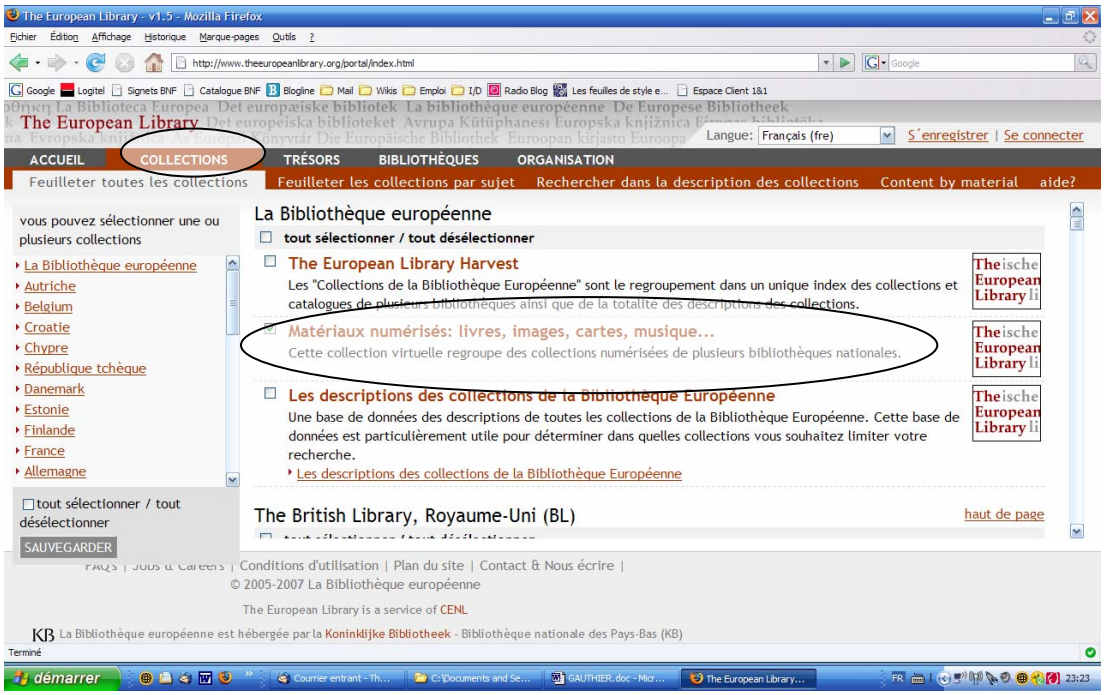
Il est plutôt rapide sur les quatre portails. Le plus court chemin est réalisé sur TEL et Europeana pour les notices BNF (4 clics), le plus long sur Europeana pour les notices BNP (6 clics) ➔ *cf Modélisation du chemin d'accès par portail en annexe pp. 115, 120, 124 et 129.*

Ce point est important dans le développement d'un portail en OAI-PMH car, dans la consultation de ressources numériques, l'utilisateur doit sortir du portail pour atteindre la ressource ; en effet, le protocole n'organise que l'échange des métadonnées. Par conséquent, trois points sont essentiels :

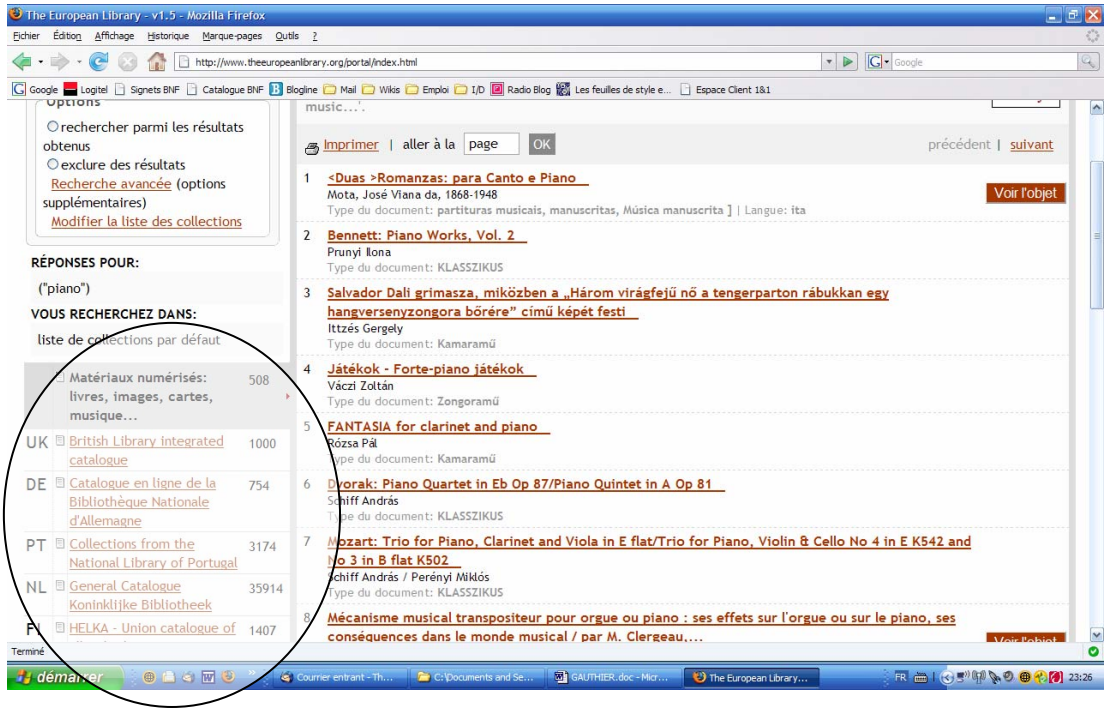
- **La possibilité de ne rechercher que des ressources numériques**

OAIster et Europeana ne propose que des ressources numériques (c'est clairement stipulé sur leur site). Sheet Music Consortium permet de ne rechercher que dans les ressources numériques (il faut cocher la case *Digitized Sheet Music Only* dans le formulaire de recherche, *cf Formulaires de recherche en annexe p. 134*).

Le cas de TEL est plus problématique. Le formulaire de recherche ne permet pas de choisir d'effectuer une recherche dans le contenu numérique seulement (il y a bien la possibilité de cocher le type de ressources *livres numérisés*, qui se noie dans des catégories aussi diverses que *article scientifique*, *liste de collections par défaut*, *cartes et atlas, cartographie* ou *musique*, ce qui est quelque peu perturbant). Après une appropriation assez longue de la façon de naviguer dans le portail, l'utilisateur comprend qu'il peut cliquer sur l'onglet *Collections*, sélectionner *Matériaux numérisés* puis retourner sur le formulaire de recherche. Sinon, lors d'une recherche dans les *Collections par défaut* (ressources numériques et notices bibliographiques), l'encart gauche de la page de résultats permet de trier selon la catégorie *Matériaux numérisés* ou par *Collection* (avec le pays en indication). Ce paragraphe qui peut paraître confus reflète un peu l'état dans lequel se trouve l'utilisateur face à TEL pour la première fois.



Onglet Collections, *choix de ne rechercher que dans les matériaux numérisés*

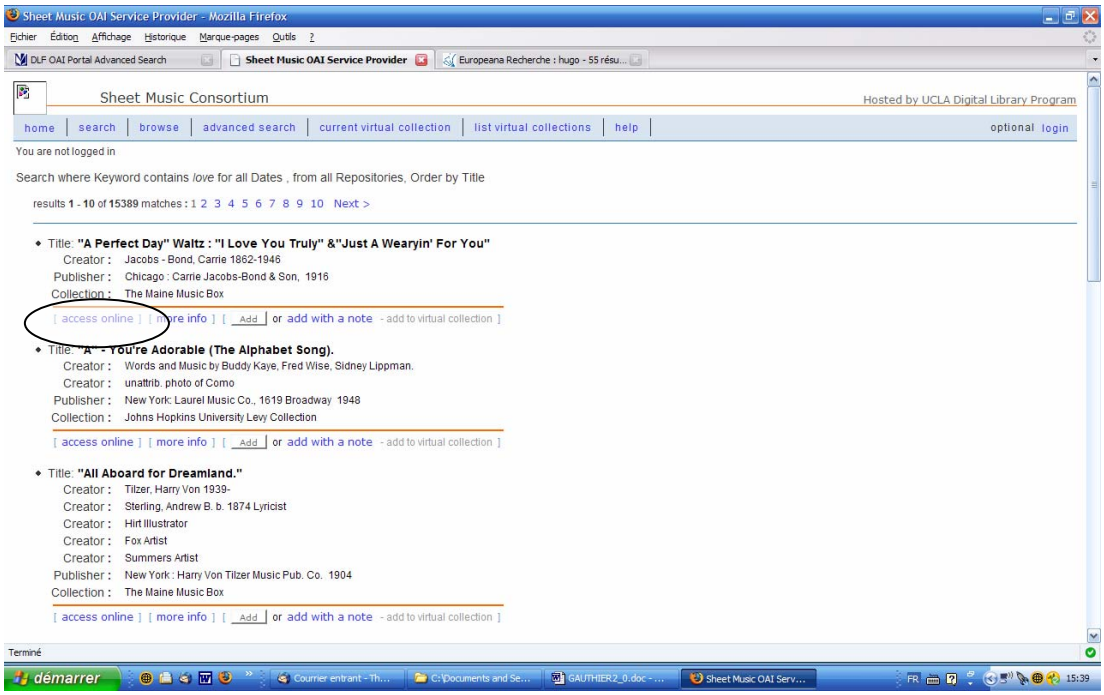


Lors d'une recherche par défaut, possibilité de trier selon les ressources numériques et les collections

- **La distinction des notices bibliographiques simples et de celles qui ont un lien à une ressource numérique**

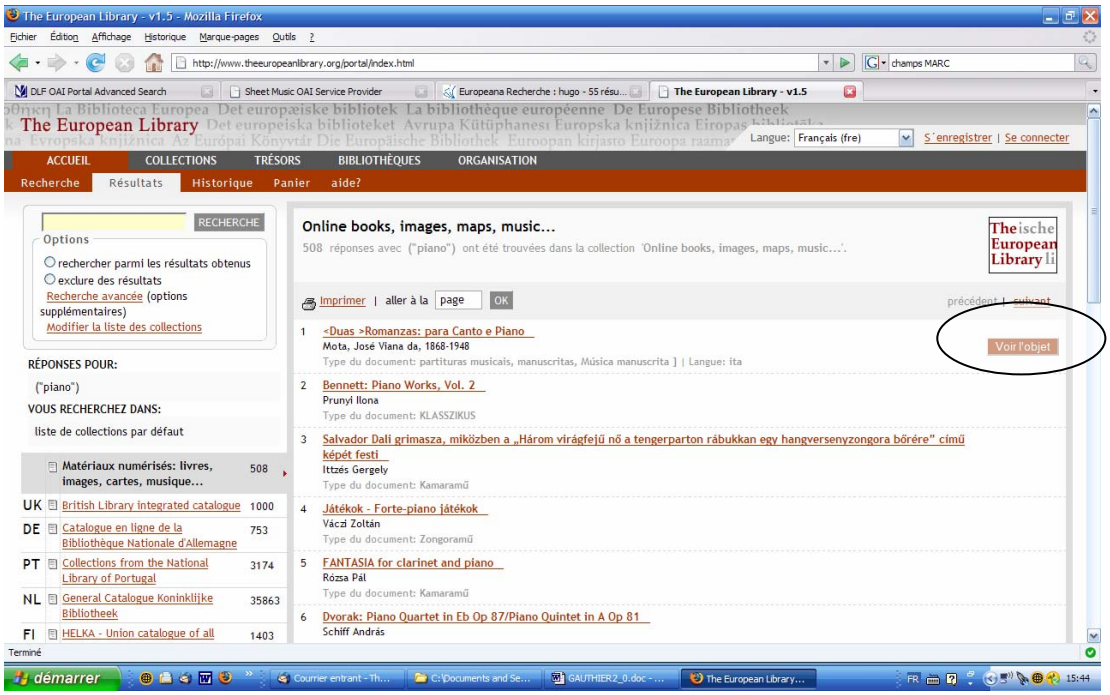
Une fois la recherche effectuée parmi les notices bibliographiques et celles qui présentent un lien à une ressource numérique, il est important que l'utilisateur distingue ces deux groupes d'un coup d'œil. Europeana et OAIster ne sont pas concernés par cette partie vu qu'ils ne proposent que des ressources numériques.

Sur Sheet Music Consortium, lors de l'affichage des résultats, les notices avec documents numériques sont identifiées clairement par un lien entre crochets *[access online]*. En cliquant, on ouvre une nouvelle fenêtre ou un autre onglet et on arrive directement sur la ressource.



Lien vers la ressource numérique

Sur TEL, là encore, c'est plus problématique. Lors de l'affichage des résultats sous forme de liste, un bouton rouge « Voir l'objet » permet de distinguer les notices avec documents numériques. Le problème, c'est que l'onglet qui permet de n'accéder qu'aux matériaux numérisés ne propose pas que des ressources numériques… et que la mention « Voir l'objet » ne garantit pas l'accès à une ressource numérique (cela peut mener à la notice bibliographique sur le site propriétaire). Cliquer sur le lien « Voir l'objet » ouvre une nouvelle fenêtre de format réduit.



Lien vers la ressource numérique sur TEL

En règle générale, selon le FD, on accède directement à la ressource ou on passe par la notice bibliographique source (sur le site propriétaire) en premier lieu, ce qui peut être gênant au niveau du confort dans la navigation.

- **Les droits d'accès relatifs aux ressources numériques décrites** (indiqués dans l'élément *rights* ou *droits*)

C'est un point essentiel dans la constitution et la consultation d'un portail qui centralise des notices provenant d'organismes propriétaires divers. Le protocole s'arrête à la notice bibliographique. L'accès au document numérique se fait par un lien qui renvoie au site propriétaire de la ressource. La question des droits de diffusion et d'accès reste donc entièrement sous la responsabilité du FD. L'élément DC *rights* permet d'indiquer quel type de droit est appliqué. Europeana ne proposant que des ressources du domaine public, aucune mention des droits n'est faite à part pour les ressources de la BNF, qui portent la mention *domaine public*. L'utilisateur n'est pas frustré car il accède toujours à une ressource numérique.

Pour les trois autres portails, l'élément *rights* est présent de façon aléatoire, selon le FD ; en outre, ce qui pose problème, c'est la valeur du champ, qui peut être ambiguë, complexe ou très longue. Voici une illustration de ce propos.

A noter : seul TEL, dans son guide des bonnes pratiques, recommande par souci de clarté de remplir ce champ à partir de la liste suivante : Free, IP based, Password required, TEL users, Pay per view.

➔ **TEL** : en effectuant la recherche simple « piano », on trouve dans les matériaux numérisés divers cas de figure : mention *domaine public* ou *free*, notice dépourvue de l'élément droit.

➔ **OAIster** : en effectuant la recherche « piano » dans les images, on trouve : pas de mention de droit (cela mène souvent à des ressources consultables), des mentions longues ou ambiguës telles que *From the collection of the History and Archives Division, Arizona State Library, Archives and Public Records. Copyright and/or publication rights for all photographs in this collection are retained by this institution*, qui n'indique pas si la ressource est consultable en ligne, des adresses de contact pour connaître la nature des droits. On voit bien ici que l'élément *droit* est finalement ambigu car une ressource peut être restreinte au niveau des droits d'exploitation mais tout de même consultable sur le site propriétaire. Dans une démarche orientée utilisateur, il y aurait plutôt intérêt à indiquer, comme le préconise TEL, si la ressource est consultable ou pas de façon très simple et très claire pour éviter toute frustration lors de la consultation. Et si l'utilisateur désire utiliser une ressource, il peut se rendre directement sur le site propriétaire et obtenir le renseignement.

D'après cette étude, il ressort que plus le portail est de taille réduite, plus son **thème** est **balisé**, plus il réunit des acteurs du même domaine, meilleure est la recherche. Il apparaît logique que le Sheet Music Consortium, qui ne propose qu'un type de ressources et qui n'agrège que six catalogues différents, ait plus de facilité à proposer des fonctionnalités qui permettent une recherche précise et transversale. En effet, même si le champ *sujet* est indexé de façon libre par les fournisseurs de données, l'homogénéité du type de ressources fait que cela ne pose pas trop de problèmes à la recherche. En revanche, les portails à **vocation encyclopédique**, dont la recherche porte sur des millions de notices disparates et appartenant à des centaines de fournisseurs de données différents ne proposent pas une recherche très efficace : même si les points d'accès sont nombreux et qu'il est possible de les croiser, la consultation des résultats est rendue difficile par le **bruit** et une très forte **hétérogénéité** des notices, tant dans le nombre des champs apparaissant que dans les valeurs assignées à ces champs. Un portail comme OAIster est conscient de ces faiblesses et travaille sur des prototypes lui permettant d'en contourner certaines. Le cluster est ainsi une façon de catégoriser a posteriori et de façon semi-automatisée un très grand nombre de notices.

Il faut noter que ces portails à vocation encyclopédique, universelle, représente tout de même un progrès dans la recherche que l'on peut faire sur le web. En effet, ils permettent une recherche sur un **fonds sélectionné** (même s'il y a un nombre important de fournisseurs de données, une sélection est opérée dans le type et la provenance des métadonnées moissonnées) et sur des **données structurées**, ce que des moteurs type Google ne proposent pas (encore ?).

Il est possible de schématiser la situation comme suit :

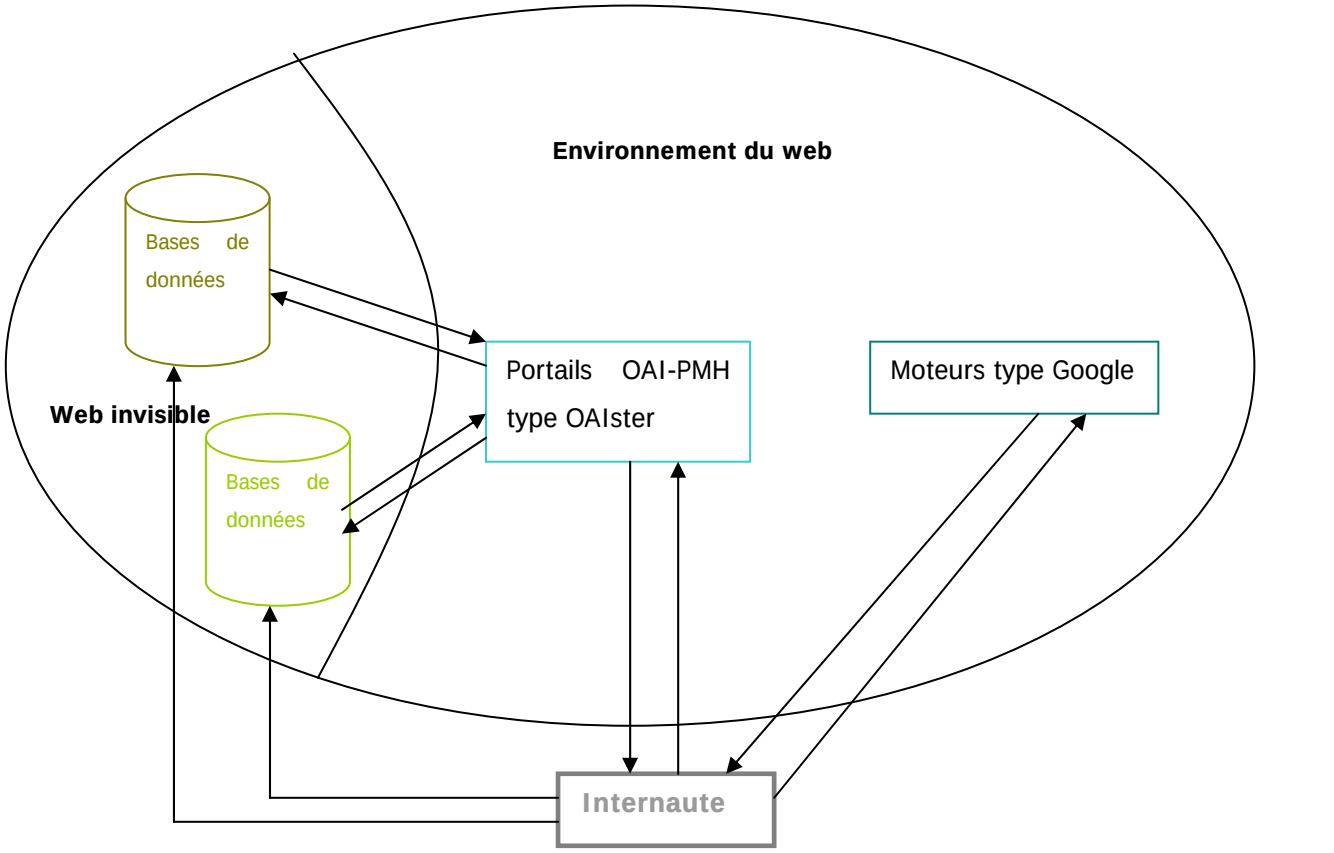


Fig. 2 : L'OAI-PMH dans l'environnement web

On peut noter également que l'OAI-PMH sert **divers objectifs**, à **diverses échelles** : dans le cas de Sheet Music Consortium ou du projet Musique contemporaine, le protocole est utilisé comme un outil permettant un **déploiement en local**, dans un domaine bien défini

alors que les projets type OAIster et The European Library ont d'une certaine façon vocation à **structurer le web** et fournir un accès simplifié à des ressources peu visibles.

Sur une note plus générale, cette tentative de catégoriser, d'organiser la nébuleuse d'informations que représente le web rentre dans une problématique très actuelle : celle du **web sémantique***. D'après Tim Berners-Lee, le premier à avoir conceptualisé cette notion, le web sémantique est une extension du web courant qui permet de structurer l'information et de lui donner un sens bien défini[1]. Le web actuel n'est pas un web « intelligent » : les moteurs de recherche « lisent » le code source des pages web mais n'ont aucun langage, aucun code pour interpréter l'information recueillie. Ils indexent les mots de la page et les classent en leur attribuant un poids (ou degré d'importance dans le document) ; ainsi, un terme que l'on retrouve dans la balise *titre* et de façon répétée dans le corps du texte aura un poids important. Au sein d'un web dit sémantique, chaque information possède une étiquette (ou *métadonnée*, une donnée sur la donnée) qui la définit, la contextualise. C'est déjà ce que fait la balise *titre* en indiquant que l'information qui y est encapsulée est un titre.

D'une certaine façon, les technologies et méthodes d'organisation de l'information développées au sein du monde des bibliothèques imprègnent les tentatives d'amélioration de la recherche sur le web. Les standards type XML et RDF (Resource Description Format) et dans une certaine mesure le protocole OAI-PMH et le format DC participent du développement du web sémantique aujourd'hui.

[1] BERNERS-LEE Tim [et al]. The semantic web. Scientific American.com [en ligne]. Mai 2001 <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>

Troisième partie
Préconisations à l'usage des fournisseurs de données et des fournisseurs de services

La littérature sur le sujet, présentée sous forme d'articles paraissant dans des périodiques spécialisés en sciences de l'information et de publications sur le web de bonnes pratiques émanant d'organismes de normalisation (type DCMI) ou d'institutions impliquées dans des projets OAI-PMH, est centrée sur l'idée que la **qualité des métadonnées** permet aux FS d'assurer une recherche précise et satisfaisante et d'offrir des services à valeur ajoutée.

L'agrégation de métadonnées provenant de bases diverses pour donner un accès centralisé à des ressources numériques est une réalité aujourd'hui, surtout depuis l'apparition du protocole OAI-PMH. Malheureusement, le résultat en termes d'interopérabilités sémantique et syntaxique et par là même, en termes de recherche, est souvent très en deçà de ce que les catalogues offrent au niveau de chaque fournisseur de données. C'est pourquoi il est nécessaire de développer des **métadonnées partageables** (« *shareable metadata* ») afin de permettre une utilisation et une compréhension des métadonnées hors de leur environnement d'origine.

Selon Liu (32, Liu), la différence essentielle entre les moteurs de recherche tels que nous les connaissons sur le web et les moissonneurs OAI-PMH tient dans le fait que le moissonneur reconnaît les formats de description et peut donc utiliser des informations structurées, permettant ainsi au fournisseur de services de proposer à l'utilisateur une recherche plus précise, plus pointue. De la qualité, de la « partageabilité » de ces métadonnées dépend donc la qualité de la recherche proposée par les FS et agrégateurs sur le portail. *Ne seront pas distingués dans cette partie l'agrégateur du fournisseur de services, la tâche de normalisation des métadonnées moissonnées leur incombant à tous deux.*

Shreeves et al. (20, 21, Shreeves) tentent de définir ce que sont des métadonnées partageables. Tout d'abord, ce sont des métadonnées qui diffèrent des métadonnées d'origine. En effet, le **contexte d'utilisation** des métadonnées est un facteur essentiel à prendre en compte pour présenter des métadonnées de qualité (ce point sera développé tout au long de cette partie). Cela implique que des métadonnées peuvent être de qualité dans un certain contexte et non dans un autre car non adaptées au nouvel environnement. Les métadonnées doivent également être **compréhensibles en soi** aussi bien par l'humain que par la machine.

Cette idée des métadonnées comme **élément central** garant d'une agrégation efficace conditionne le contenu des bonnes pratiques trouvées. L'idée largement répandue que l'OAI-PMH est facile à mettre en place car il suffirait aux FD de mettre à disposition leurs données dans un entrepôt est inexacte. En effet, si, techniquement, l'OAI-PMH est assez simple à implémenter, le développement auquel procède le FS est rendu d'autant plus difficile que les FD proposent des métadonnées de mauvaise qualité et/ou non appropriées. Ainsi, tous les

guides consultés sur le sujet non seulement distinguent les bonnes pratiques à l'attention du FD de celles à l'attention du FS mais en plus insistent sur le travail important que le FD doit mener en amont sur ses métadonnées.

Le sujet du mémoire, la nature de la discipline dans laquelle il s'insère ainsi que la nature souvent très technique des tâches effectuées par le fournisseur de services dans la mise en place d'un portail s'appuyant sur l'OAI-PMH font que les préconisations à l'usage des FS sont moins importantes en quantité que celles à l'attention des FD.

1 Préconisations à l'usage des fournisseurs de données

Shreeves a résumé dans l'article *Moving towards shareable metadata* (21, Shreeves) les six caractéristiques qui définissent des métadonnées « partageables ».

Les six caractéristiques pour des métadonnées partageables ou les six « C » :

- **Contenu** :
    - La granularité doit être appropriée. Elle se situe souvent au niveau du document-objet mais il est parfois plus pertinent de décrire la collection ; tout dépend du projet dans lequel on s'insère ;
    - Le contenu d'éléments spécifiques doit être précisé (indiquer quel vocabulaire contrôlé est utilisé dans le champ *sujet* par exemple.)
- **Cohérence** (*consistency*) : nécessité d'une cohérence sémantique et syntaxique. Cela facilite le travail du FS (un manque de cohérence dans le nombre d'éléments utilisés ou dans la valeur donnée à un type d'élément n'aide pas le FS dans l'affichage des notices et l'indexation des champs). Idem pour les vocabulaires contrôlés et les normes appliquées. La clef pour le FS est la **prévisibilité**, ce afin de pouvoir appliquer des techniques d'homogénéisation à de larges groupes de métadonnées.
- **Cohérence** (*coherence,* avec l'idée de clarté, d'intelligibilité) : les métadonnées doivent être **compréhensibles en elles-mêmes**. Les éléments et les valeurs doivent être associés logiquement, il doit y avoir une information par élément (par exemple, l'élément *titre* ne doit pas contenir l'auteur après un slash). Au besoin, il est possible de répéter l'élément.
- **Contexte** :
    - L'introduction d'information sur le contexte de création et d'utilisation des métadonnées est souvent nécessaire ;
    - L'inverse est également vrai : certaines données essentielles en local ne servent à rien dans un environnement partagé.

- **Communication** :
    - Les FS peuvent utiliser ce qu'ils savent de la façon dont les FD ont créé leur métadonnées ;
    - A l'inverse, savoir comment le FS opère peut aider les FD à proposer des métadonnées de meilleure qualité.
- **Conformité à des standards** : le FD doit se conformer à des normes et standards reconnus (OAI-PMH, XML, DC).

Ces caractéristiques offrent une bonne introduction aux préconisations à l'usage des fournisseurs de données.

1.1 Préconisations techniques

1.1.1 Cadre du projet (objectif, public, contexte)

Dans le cadre d'un projet, il est nécessaire de **définir un objectif** et d'**identifier le public-cible** pour pouvoir adapter ses métadonnées à l'usage qu'il en sera fait dans le portail.

> ➔ Dans le cadre d'un portail qui présente des ressources pédagogiques à destination de professeurs et d'élèves, l'indication du type d'audience pour chaque ressource devient essentielle.

L'objectif et le public étant identifiés, il est important d'**adapter ses métadonnées** au nouvel environnement en ajoutant des éléments permettant de les comprendre dans leur nouveau contexte (7, Prom ; 18, Hutt ; 20, 21, Shreeves ; 29, Hillmann).

> ➔ Dans un nouvel environnement, il peut être nécessaire de rajouter le nom de la collection à laquelle appartient une ressource. L'exemple de Wendler (cité dans 21, Shreeves) appelé *On a horse* est particulièrement explicite : une collection de photos de Theodore Roosevelt apparaît dans un portail agrégé. Une des photos ne porte pour titre que la mention *On a horse*. Dans le contexte d'origine, cela fait sens car toute la collection est à propos de Theodore Roosevelt ; en revanche, dans le nouvel environnement, la mention de l'appartenance de la photo à la collection Theodore Roosevelt devient indispensable pour comprendre la notice.

A l'inverse, certaines informations, inutiles dans le nouveau contexte d'exploitation, peuvent être retirées (18, Hutt ; 20, 21, Shreeves).

> ➔ Europeana ne propose que des livres numérisés ; la présence d'un champ *type* paraît donc inutile.

1.1.2 Lots

La constitution d'entrepôts selon une **organisation réfléchie** permet au FS d'améliorer par la suite la recherche. La hiérarchisation ou la division des enregistrements en ensembles et sous-ensembles cohérents est ainsi fortement recommandée.

Il est donc parfois utile d'ajouter aux notices un champ apportant des informations sur le type ou sur l'origine de la ressource. Cela permet de créer un critère de dissociation ou de regroupement des notices d'un même FD mais aussi de FD différents. Cet ajout de champs peut être également effectué par le fournisseur de services qui a en charge le développement de la recherche à partir des notices dont il dispose.

> ➔ Comme nous l'avons vu lors de l'étude, la plupart des portails offrent la possibilité de rechercher par fournisseur de données, ce qui s'avère très pratique pour l'utilisateur. Cela est rendu possible par l'ajout d'un champ renseignant sur l'origine de la notice.

1.1.3 Formats de description

Le format Dublin Core non qualifié est requis pour assurer une interopérabilité technique minimale mais il est vivement conseillé de décrire ses métadonnées sous le **format** (encodable en XML) **le plus riche possible** (18, Hutt ; 23 ; 29, Hillmann ; 32, Liu ; 33, 34, Tennant). Plus les métadonnées seront précises, plus les fonctionnalités de recherche développées par le FS seront fines.

Le **choix du format** des métadonnées doit être adapté aux buts fixés dans le projet (pour décrire des archives, EAD est préférable à MODS par exemple) (23).

En outre, il est conseillé de suivre les **bonnes pratiques** existantes relatives à la constitution des équivalences entre formats ou *mapping**.

> ➔ Par exemple, il existe des bonnes pratiques pour passer du format EAD (Encoded Archival Description) au DC (7, Prom), tout comme le site de la DCMI propose des préconisations en termes de *mapping* entre les formats MARC et DC.

➔ Dans le passage d'un format riche au format DC non qualifié, la DLF conseille de procéder par pallier en passant d'abord par le DC qualifié (23).

➔ Le passage d'un format riche au format le plus simple qu'est le DC nécessite parfois de créer, pour une notice en format enrichi, plusieurs notices en DC (7, Prom).

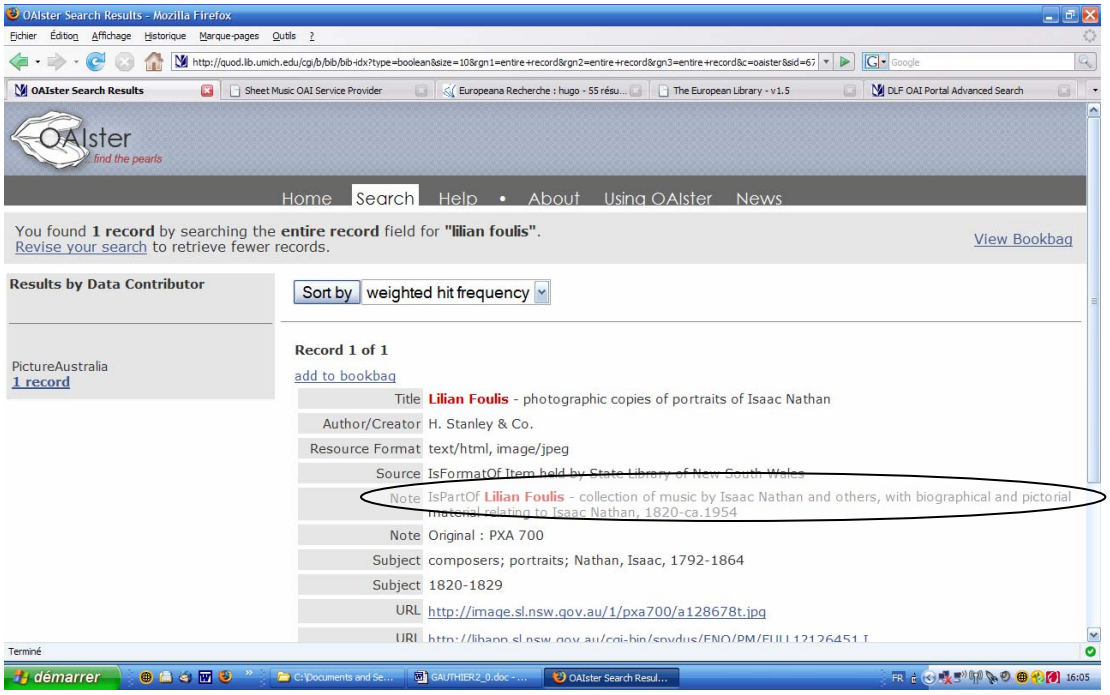
1.1.4 Granularité

Bien réfléchir en amont au **degré de granularité** le plus adapté au projet de portail. Selon le contexte, une description au niveau de la collection ou du document sera plus appropriée.

➔ Pour la DLF (23), le niveau de granularité le plus fin est à privilégier en général. En revanche, concernant certaines ressources qui n'existent que sous la forme numérique (sites web, tutoriels par exemple), la DLF préconise de rester au niveau de granularité le plus large (décrire le site web dans son ensemble et non au niveau des pages).

Il est possible d'indiquer le contexte d'une notice et de maintenir les liens qui existent entre des notices de niveaux de granularité différents.

➔ Dans le format DC qualifié, on trouve l'élément *relation* et ses qualificatifs *IsPartOf* et *HasPartOf*, qui permettent de décrire les liens.



Sur cette notice OAIster provenant de Picture Australia, le contexte est donné de façon très succincte par la mention IsPartOf *du champ* Note.

➔ Avec un format tel que MODS, la granularité ne pose pas de problème. L'arborescence de notices mères et filles est recréée en XML.

1.1.5 Métadonnées

Que ce soit dans le cadre d'un projet spécifique ou pas, certaines caractéristiques garantissent la partageabilité des métadonnées.

Les métadonnées doivent être les plus cohérentes possibles et ce pour faciliter le travail de normalisation et de développement de l'interface de recherche par le FS. La **cohérence** s'apprécie à des degrés divers :

- Tout d'abord, faire attention à ce que l'**encodage** (en XML, en DC) soit bien propre. Des erreurs d'encodage rendent plus difficile le travail du FS de rappel des notices lors de la recherche ;

- Il est préférable que le nombre de **champs** (ou éléments dans le DC) soit le même pour chaque ressource. Une grande variation au sein d'une collection pose des problèmes au FS pour le traitement et l'affichage ;

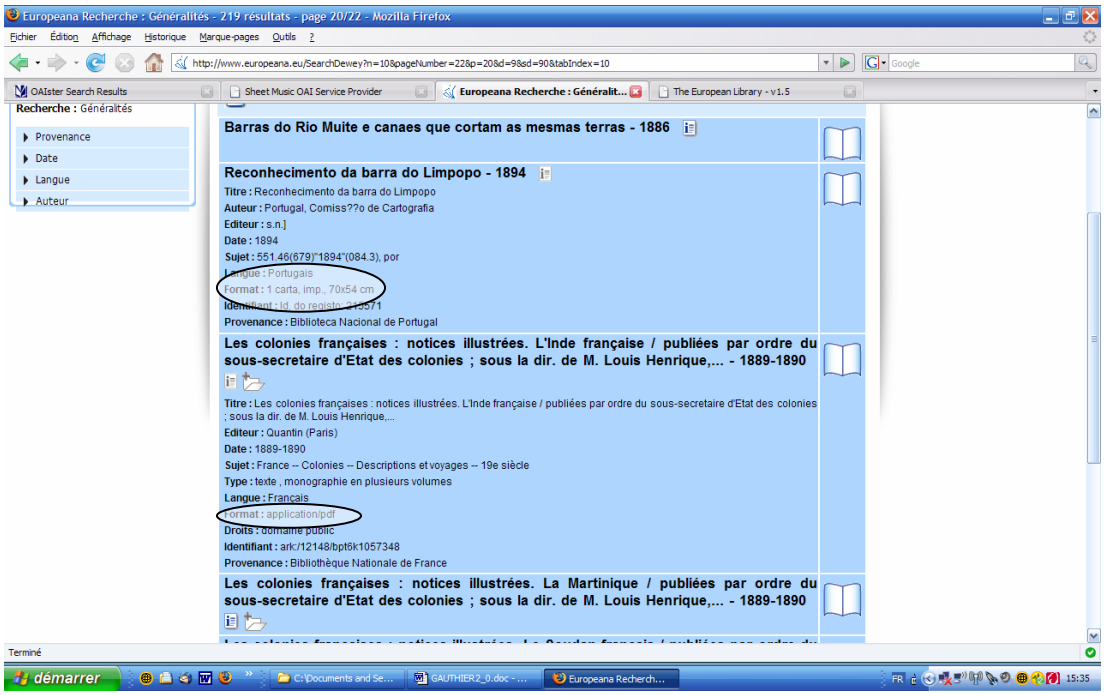
- Il doit également y avoir cohérence au niveau du **type de valeur** associée à un champ.

  ➔ Si, dans le cadre du format DC non qualifié, le champ *couverture* est utilisé, il est déconseillé d'y décrire la couverture spatiale pour certaines ressources et la couverture géographique pour d'autres.

  ➔ Idem pour le champ *format* : sur Europeana, il est renseigné soit par le type de format numérique (pdf), soit par la taille de la ressource (23 cm). Vu que seuls des livres numérisés sont présents sur ce site, la mention du format informatique semble plus intéressante car elle renseigne sur l'accessibilité au document numérique. D'ailleurs, on peut noter que ces notices ne décrivent pas la même chose : l'une décrit le document primaire, la ressource et l'autre décrit une autre forme de cette ressource (le document numérique qui en est tiré).

  Sur une note plus générale, le renseignement de ce type de champ de description apporte des informations sur la crédibilité de l'institution dont

émanent les notices. La question sous-jacente pourrait être : que donne-t-on à lire à l'utilisateur ? Accède-t-il au document primaire ou bien à une forme dérivée du document primaire ? Ces questions, non essentielles dans le cadre du mémoire, prennent tout leur sens lorsqu'on se trouve dans le contexte d'informations financières ou statistiques par exemple.



Le champ format *n'est pas renseigné de la même façon sur ces deux notices d'Europeana*

- En DC, il est conseillé de suivre le **principe du *one-to-one*** : à un élément correspond une valeur. Cela a pour but, en plus de faciliter le travail de développement du FS, de rendre la recherche sur un champ plus précise. Voici un exemple de métadonnées posant problème et ce qu'il est conseillé de faire par la DCMI :

  ➔ Creator : Schwartz, Jean ; Francis, André

  Préférer :

  Creator : Schwartz, Jean

  Creator : Francis, André

Il est préférable de répéter le champ autant de fois que nécessaire plutôt que de remplir d'informations un même champ qui deviendra difficile à exploiter par la suite.

- L'utilisation de **normes** nationales ou internationales, de vocabulaires contrôlés, qu'ils soient le fait d'organismes de normalisation ou propres à une institution, est vivement conseillée. Cela permet d'avoir des métadonnées les plus homogènes possibles (17, 29, Hillmann ; 18, Hutt ; 21, Shreeves ; 33, 34, Tennant).

  ➔ Le champ *date* est particulièrement représentatif de ce problème. Si le champ *date* est encodé des façons suivantes dans une unité de notices qui fait sens :

  1945

  1990-1995

  1995-02-03

  le travail de normalisation pour le FS sera complexe et important et la recherche par date n'en sera que plus hasardeuse.

  Il est donc conseillé de choisir une norme d'écriture (ISO8601 par exemple) et de s'y tenir.

- Il est conseillé d'indiquer quels vocabulaires contrôlés et quelles normes sont appliqués, grâce aux qualificatifs du DC qualifié (d'autres formats permettent d'indiquer cette information également) ;

- La NSDL conseille d'utiliser pour le champ *sujet* différents vocabulaires contrôlés si cela est possible. Cela donne une plus grande latitude au FS pour améliorer les possibilités de recherche transversale ;

Trois indexations différentes sur la DLF : la première est l'indexation DC originale, la deuxième et la troisième proposent des sous-domaines et des domaines, tous deux issus du clustering.

- S'il existe des bonnes pratiques pour encoder les métadonnées dans un format (Dublin Core, EAD,...) particulier, il est souhaitable de les suivre (7, Prom ; 18, Hutt). Pour cela, il est nécessaire de rechercher les normes, les standards, les guides de bonnes pratiques auprès des organismes et institutions professionnelles les éditant. Par exemple, l'encodage en EAD est expliqué sur le site des Archives de France[1], celui en Dublin Core, sur le site de la DCMI[2] (Dublin Core Metadata Initiative).

1.1.6 Droits

Les préconisations de TEL (24) en termes de droits semblent intéressantes en ce qu'elles sont **orientées utilisateur**. En effet, il est essentiel pour l'utilisateur d'un portail de savoir s'il peut ou non accéder à la ressource. Remplir le champ *droit* d'informations juridiques complexes à décrypter paraît contreproductif et peut faire naître une frustration chez l'utilisateur, qui se trouvera déçu de n'avoir pas compris que l'accès à une ressource était limité. En résumé, le maître-mot dans ce domaine pourrait être **prévisibilité** ; l'utilisateur

[1] http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/DAFlangage.html
[2] http://dublincore.org/usage/terms/

doit pouvoir distinguer le plus tôt possible ce qui est consultable en ligne de ce qui ne l'est pas.

➔ Il est par exemple possible d'ajouter un champ *accès à la ressource numérique* indiquant les modalités de consultation du document par *oui*, par *non* ou bien sous certaines conditions (abonnement, paiement, etc…) Il est également possible de matérialiser cet accès par des icônes très simples, par des couleurs (le rouge, le vert et l'orange que l'on retrouve sur les feux de circulation), etc…

1.2 Préconisations organisationnelles

La DLF (23 ; 29, Hillmann) préconise de fournir de la **documentation** au FS sur les points suivants :

- Historique de la création des métadonnées ;
- Format d'origine si différent de DC ;
- Explicitation des choix effectués en matière de format et de métadonnées ;
- Tableau d'équivalence entre les formats ;
- Utilisation de vocabulaires contrôlés (indiquer lesquels, concernant quels éléments ; cela est particulièrement important quand le DC seul est utilisé) ;
- Signification des dates (surtout quand seul le DC est utilisé) ;
- Procédures de mises à jour (quand, comment) ;
- Lots (lesquels).

Ces informations peuvent être fournies en partie **en interne** (dans les descriptifs des entrepôts, dans les descriptifs des sets, dans les enregistrements) et en partie **en externe** (par la publication de pages web par exemple).

Le travail de normalisation et de préparation des métadonnées à la collecte à partir d'un entrepôt est facilité par une **organisation** et une **communication** efficaces.

➔ La mise en place de réunions régulières et bien balisées non seulement en amont du projet mais aussi tout au long du processus, la constitution de groupes de travail se partageant les tâches, l'utilisation d'outils aidant au travail collaboratif (wikis, mail, mise

à disposition des documents de travail sur un serveur, etc…), la nomination d'un garant des bonnes pratiques sont autant de façons de coordonner efficacement le travail en mode projet.

2 Préconisations à l'usage des fournisseurs de services

2.1 Préconisations techniques

Ce sont les services informatiques des fournisseurs de services pilotes des projets en OAI-PMH qui vont répondre aux questions techniques. Ils vont devoir entre autre travailler sur les métadonnées agrégées, mettre en place la recherche à partir de critères communs aux notices. Derrière cet aspect technique se trouve des questions purement documentaires auxquelles le documentaliste doit réfléchir en amont du projet. Il est donc nécessaire que les services informatique et de documentation travaillent ensemble.

2.1.1 Compréhension des métadonnées

Il est nécessaire de comprendre **comment les éléments ont été interprétés** par les FD et comment ces derniers ont réalisé le passage d'un format à un autre.

Il est également conseillé d'**identifier les vocabulaires utilisés**, ce dans le but de déterminer s'il est possible d'utiliser un vocabulaire contrôlé commun à tous (26, Cole). Les enregistrements peuvent être liés a posteriori à un **thésaurus externe**. Une table d'équivalences permet alors de faire le lien entre les enregistrements et les termes du thésaurus (4, Foulonneau).

> ➔ Dans le cadre d'un projet bien organisé, les éléments nécessaires à la compréhension des métadonnées sont inclus par les fournisseurs de données, comme nous l'avons vu précédemment. Cependant, l'institution qui assure le développement technique du portail peut demander aux différents acteurs l'ajout de ce type d'informations **en amont** du projet.

2.1.2 Normalisation des métadonnées

Si le FD a intérêt à normaliser ses métadonnées le plus possible, le FS doit aussi opérer un travail de **normalisation** des données des collections moissonnées et ce afin de proposer une recherche la plus efficace possible sur le portail (18, Hutt ; 31, Liu ; 33, 34, Tennant).

Il est possible de mettre en place un **contrôle qualité** sur les données moissonnées qui permet de rejeter les enregistrements incomplets, de vérifier leur validité XML et d'éliminer les doublons. En outre, *« un **normaliseur*** (application de retraitement) peut effectuer un*

contrôle syntaxique, ajouter des valeurs et réorganiser les enregistrements » (4, Foulonneau).

2.1.3 Création de profils d'application

Le FS peut décider, selon les objectifs visés, de créer un schéma de métadonnées* en XML à partir de schémas existants. Ainsi, TEL (24) a mis en place ce que l'on appelle un profil d'application *(application profile)* basé sur des éléments DC existant et enrichi par des éléments créés par l'institution elle-même et nommés *tel*. Cela se concrétise comme suit :

➔ <dc>

<dc:title>This is a title</dc:title>

</dc>

<tel>

<tel:type>This is a map</tel:type>

</tel>

Le titre de la notice est renseigné par un champ appartenant au format Dublin Core et le champ type de la ressource ainsi que la liste contrôlée correspondante sont créés par TEL.

2.2 Préconisations organisationnelles

D'après de nombreux acteurs ayant travaillé sur des portails s'appuyant sur l'OAI-PMH, le mot d'ordre semble être la communication : communication entre acteurs d'un projet, entre fournisseurs de données et fournisseurs de services mais aussi communication et travail collaboratif entre fournisseurs de services appartenant à des projets différents afin de partager l'expérience accumulée.

- **Communication avec les FD**

  Le FS n'a pas intérêt à accepter des métadonnées non « traitées », non adaptées au nouvel environnement. Il est conseillé de fournir aux FD des guides de bonnes pratiques mettant l'accent sur l'importance de suivre des normes et des vocabulaires contrôlés (11, Lagoze ; 33, 34, Tennant).

➔ Cela peut se concrétiser par l'édition et la publication de **guides** papier, de guides en ligne sur un site web ou sur un wiki (ce qui permet la mise en commun des connaissances). Un **coordinateur**, qui assure la centralisation et la transmission de l'information entre le fournisseur de services et les fournisseurs de données peut également être nommé.

- **Communautés de pratiques**

  Il est conseillé d'entrer en contact avec d'autres FS pour échanger ou travailler ensemble sur des processus de normalisation automatique des métadonnées, sur le développement d'algorithmes de recherche. Pour cela, il faut rechercher dans son domaine professionnel (par l'intermédiaire d'associations professionnelles, de colloques, etc…) les acteurs susceptibles d'être intéressés par une collaboration sur ce type de projets.

2.3 Limites à ces préconisations et solutions proposées

Dans le cadre d'un projet à très grande échelle (OAIster, NSDL), les préconisations en termes de normalisation des métadonnées émanant du FS et en direction des FD ont leur limite ; il semble en effet impossible d'imposer à un très grand nombre de FD un traitement fin de leurs métadonnées. **C'est donc le FS qui a la charge de traiter les métadonnées moissonnées en parant à un grand nombre d'erreurs et d'approximations**.

Ainsi est née au sein de la NSDL une technique nommée *safe transform* (que l'on pourrait traduire par *transformation sécurisée*) qui vise à modifier les métadonnées en XML de divers FD (30, Hillmann). Ce processus a pour but de mettre en valeur l'information contenue dans les métadonnées sans en altérer la teneur. Il se divise en trois catégories :

- **Limiter le bruit**, en retirant des valeurs inutiles (ponctuation, valeur type « inconnu », « s.d. ») ;
- **Identifier les vocabulaires contrôlés** utilisés pour pouvoir en tirer partie. Cela fonctionne pour les vocabulaires limités du genre DCMI Type ;
- **Normaliser la présentation des métadonnées**, en « nettoyant » les valeurs, en retirant les encodages XML en double,…

Cependant, certaines modifications ne peuvent se faire qu'au niveau de chaque collection, rendant la tâche plus ardue.

L'article fait part d'une autre méthode de mise en valeur des métadonnées, cette fois au niveau de l'élément et non plus de l'enregistrement. Il s'agit d'utiliser les métadonnées incomplètes d'une même ressource mais de provenance diverse pour **créer une notice complète**. Ainsi, Zeng (35, Zeng) donne un exemple provenant de la NSDL. Une nouvelle notice OAI-PMH est créée à partir de trois FD : le premier fournit les éléments DC *titre*, *identifiant*, *createur* et *type* ; le second donne des informations quant au public et au niveau d'études ; le troisième a indexé la ressource à l'aide de trois vocabulaires contrôlés. Cette méthode permet de **retirer les doublons** et de proposer une **notice enrichie**.

De même, OAIster ne peut exiger un traitement fin des métadonnées de ses quelque 890 fournisseurs de données (au 8 octobre 2007) ! C'est pourquoi ses développeurs ont mis en place des techniques semi-automatisées de traitement des métadonnées moissonnées, tel que le **cluster** abordé en deuxième partie. Le cluster consiste à « *retenir les mots et les expressions qui constituent les métadonnées et à les regrouper en groupes sémantiques* » (28, Hagedorn). Chez OAIster, cette catégorisation ne se fait pas sur tous les mots des champs ; l'algorithme utilisé permet de créer automatiquement de très larges catégories en identifiant les mots dans leur contexte syntaxique. On se retrouve alors avec des séries de mots regroupés sous des étiquettes ; certaines sont pertinentes (par exemple *church*, *religious*, *christ*, *bishop* sous *christianity*), d'autres ne le sont pas et sont donc rejetées. Puis les étiquettes sont organisées selon un **système de classification** existant. Enfin, les catégories et sous-catégories sont incluses dans les métadonnées (28, Hagedorn). Cependant, la technique du clustering produit encore beaucoup d'imprécisions et d'inexactitudes dans l'indexation effectuée a posteriori.

Ce même portail est en train de développer un projet appelé DLF MODS[1] tout entier tourné vers la mise en place de méthodes pour transformer et améliorer la recherche de ressources décrites au format MODS (Metadata Object Description Schema). MODS est un format XML permettant de décrire des ressources au format MARC 21. Il permet de faire un pont entre un format de description riche, MARC 21, et un format simple, Dublin Core (35, Zeng).

On peut également émettre des limites à propos de la mise en place de communautés de pratiques, assez développées dans le domaine de la culture mais qui peuvent se révéler beaucoup plus problématiques dans le secteur capitalistique, où la concurrence amène plutôt ses acteurs à garder l'information et le retour d'expérience pour soi.

[1] http://quod.lib.umich.edu/m/mods/

3 Tableau récapitulatif des préconisations à l'usage des fournisseurs de données et des fournisseurs de services

Ce tableau récapitulatif se veut une sorte de « checklist » à l'usage des acteurs désireux de développer un projet en OAI-PMH. La responsabilité de la tâche est indiquée pour chaque préconisation.

Type de préconisations	**Préconisations**	**Fournisseurs de données**	**Fournisseurs de services**
Technique	Respect du protocole OAI-PMH (langage XML, encodage en DC non qualifié)	X	
	Choix du format le plus riche pour la description de données	X	
	Granularité	X	X
Pour assurer une interopérabilité syntaxique et sémantique	Bonnes pratiques concernant l'utilisation du format DC ou autres	X	
	Vocabulaires contrôlés et normes	X	
	Normalisation automatisée en aval des métadonnées moissonnées		X
Organisationnel	Communication sur projet	X	X
	Communication sur objectifs du portail		X
	Communication sur métadonnées	X	
	Echange d'expérience avec homologues	X	X

Tab. 3 : Préconisations générales

Si chaque type d'acteurs se voit confier des tâches bien distinctes, la préconisation qui concerne de façon égale les deux types de fournisseurs est la **communication**. A tous les stades du projet, FD et FS doivent échanger des informations qui permettent de **comprendre ce que chacun fait**. Ainsi, le FS rend la tâche de préparation de ses métadonnées par le FD plus facile s'il explique ce dont il a besoin pour obtenir une recherche la plus efficace possible. De même, le FD, en expliquant dans les données elles-mêmes mais aussi à l'aide d'une documentation complémentaire quels formats sont utilisés, comment les équivalences entre eux ont été décidées et réalisées, quels vocabulaires contrôlés ils utilisent, fera gagner du temps au FS (18, Hutt ; 33, 34 Tennant).

Conclusion

L'OAI-PMH crée un standard technique d'échange de métadonnées. Cependant, cette interopérabilité technique est nécessaire mais non suffisante pour réaliser des portails à la recherche transversale, qui est le but sous-jacent de tout portail OAI-PMH. Aussi, le véritable enjeu ne réside-t-il pas dans la mise en place d'une interopérabilité syntaxique mais aussi sémantique et organisationnelle pour rendre un peu plus homogène ce qui au départ ne l'est pas ? Au fond, ce travail sur l'interopérabilité, réalisé en partie par les services informatiques pour ce qui est de la technique, nécessite des compétences documentaires et peut permettre au documentaliste de se repositionner dans son entreprise.

En outre, la communication et la mutualisation, matérialisées par l'échange autour de l'expérience, par la rédaction de préconisations et de bonnes pratiques, par la mise en place de langages communs, trouve une place prépondérante dans la construction de portails intégrés. On voit d'ailleurs se développer sur le sujet blogs, wikis, pages web de bonnes pratiques et autres avatars de ce que l'on appelle désormais le web 2.0.

Cela dit, tous les portails OAI-PMH ne remplissant pas les mêmes fonctions, les préconisations doivent être adaptées aux buts fixés. Comme on l'a vu, les objectifs du portail conditionnent le type de travail à réaliser pour obtenir une recherche efficace. Ainsi, les portails de petite taille, dont le projet est balisé et dont les acteurs appartiennent à la même communauté professionnelle réussissent plus aisément à mettre en place une recherche intéressante et un affichage homogène. Ceux-là possèdent des métadonnées souvent très travaillées et répondant déjà à des normes de description. L'OAI-PMH est pour eux un outil dont ils se servent pour se rendre plus visibles et pour développer des projets et des partenariats nouveaux à l'intérieur de leur domaine professionnel. En revanche, les gros portails à vocation encyclopédique, dont les acteurs sont nombreux et très divers dans leurs pratiques professionnelles proposent une recherche qui fait énormément de bruit et dont les résultats sont difficiles à exploiter. On voit ici se dessiner deux façons d'envisager l'OAI-PMH : pour les premiers, c'est tout simplement un nouvel outil qu'ils adaptent à leur environnement particulier ; pour les autres, c'est un premier pas vers un accès à des ressources innombrables mais mieux structurées et plus fiables.

L'OAI-PMH se trouve à mon sens à la croisée des chemins entre des bases de données hyper-structurées mais peu visibles et le web, très riche en ressources mais pauvre au niveau de la recherche. Et certains travaux effectués au sein de projet OAI-PMH (chez OAIster notamment) apportent un début de réponse (cluster, utilisation de thésaurus et d'ontologies, normalisation à grande échelle) à certaines questions posées dans le cadre du web sémantique. Finalement, l'OAI-PMH s'insère bien dans les problématiques émergentes du web sémantique et c'est peut-être aussi pour cette raison qu'il est en développement.

Bibliographie

Bibliographie analytique arrêtée au 24 août 2007. La rédaction des références bibliographiques est conforme aux normes :

- **Z44-005 décembre 1987**, *Information et documentation. Références bibliographiques. Contenu, forme et structure* ;
- **NF ISO 690-2 février 1998**, *Information et documentation. Références bibliographiques. Partie 2, Documents électroniques, documents complets ou parties de documents*.

Les notices sont classées de la façon suivante :

Cette classification suit la progression à l'œuvre dans le mémoire. A l'intérieur de chaque rubrique, les notices sont classées par ordre alphabétique d'auteur puis par ordre alphabétique de titre lorsqu'il n'y a pas d'auteur identifié.

L'OAI-PMH

Généralités

[1] The Open Archives Initiative Protocol for Metadata Harvesting. Open Archives [en ligne], 2002, 12 oct. 2004, [consulté le 22 sept. 2007]. <http://www.openarchives.org/OAI/openarchivesprotocol.html>

Version 2.0 du protocole OAI-PMH, ce document définit les concepts et les aspects techniques du protocole et propose un guide d'implémentation et un glossaire.

L'OAI-PMH et le domaine patrimonial

[2] ARMS Caroline R. Available and Useful: OAI at the Library of Congress. Library Hi Tech [en ligne], 2003. [consulté le 10 août 2007], Vol. 21, n° 2, p. 129-139. <http://memory.loc.gov/ammem/techdocs/libht2003.html>

Après une courte définition du protocole OAI-PMH, l'article revient sur l'utilisation précoce de l'OAI-PMH au sein de la Bibliothèque du Congrès.

[3] BOSTON Tony. Exposing the deep web to increase access to library collections. National Library of Australia [en ligne], 2005. [consulté le 10 août 2007], <http://www.nla.gov.au/nla/staffpaper/2005/boston2.html>

La bibliothèque nationale d'Australie n'échappe pas à la problématique de la visibilité sur le web. Elle a donc développé des portails utilisant l'OAI-PMH pour exposer certaines de ses ressources. L'article est intéressant en ce qu'il donne une bonne définition du web invisible.

[4] FOULONNEAU Muriel. Collaborer pour de nouveaux services culturels en ligne : le protocole OAI, protocole de collecte de métadonnées de l'Initiative des Archives Ouvertes. Relais Culture Europe pour la Mission de la Recherche et de la technologie du Ministère de la Culture et de la Communication et le projet européen MINERVA [en ligne], janvier 2004. [consulté le 10 août 2007], <http://www.culture.gouv.fr/mrt/numerisation/fr/technique/documents/guide_oai.pdf>

Rédigé par la spécialiste française du domaine, le document présente le protocole OAI-PMH puis développe les implications liées à la fourniture de données et à la fourniture de services. C'est un travail complet et abouti sur le protocole (première partie) et sur le protocole dans

le domaine patrimonial (deuxième partie). Il constitue une bonne approche pour ceux qui ne connaissent pas l'OAI-PMH et c'est l'un des rares documents en français.

[5] GATENBY Janifer. Aiming at quality and coverage combined: blending physical and virtual union catalogues. Online Information Review, 2002, vol. 26, n° 5, p. 326-334.

L'article compare trois façons de structurer une recherche au sein d'un catalogue : le modèle distribué (protocole Z39.50), le modèle distribué avec index centralisé (OAI-PMH) et le catalogue général. Bien utile pour comprendre ce qu'est une architecture distribuée.

[6] NAWROCKI François. Le protocole OAI et ses usages en bibliothèque. Ministère de la Culture et de la Communication [en ligne], janvier 2005. [consulté le 10 août 2007], <http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm>

Le document définit très clairement le protocole OAI-PMH et expose les raisons de son utilisation par les bibliothèques. C'est le document rédigé en français le plus facile d'accès.

[7] PROM Christopher J., HABING Thomas G. Using the Open Archives Initiative Protocols with EAD. In Proceedings of the 2nd Joint Conference on Digital Libraries [en ligne]. 14-18 juillet 2002, New York, Gary Marchionini and William Hersch, Association for Computing Machinery. [consulté le 20 août 2007] <http://dli.grainger.uiuc.edu/publications/jcdl2002/p14-prom.pdf>

L'article aborde les problèmes spécifiques posés par la mise en équivalence du format EAD (Encoded Archival Description) avec le format Dublin Core. L'une des solutions consisterait à « éclater » un enregistrement EAD en plusieurs enregistrements DC.

[8] SHREEVES Sarah L. [et al]. Current development and future trends for the OAI Protocol for Metadata Harvesting. Library Trends, printemps 2005, vol. 53, n° 4, p. 576-589.

L'article fait un tour d'horizon de l'utilisation de l'OAI-PMH hors du champ des pré-publications scientifiques. Il en ressort que les fournisseurs de services doivent faire face à des difficultés dans la mise en place de services à valeur ajoutée.

[9] VAN DE SOMPEL Herbert, YOUNG Jeffrey A., HICKEY Thomas B. Using the OAI-PMH... differently. D-Lib Magazine [en ligne], juillet/août 2003, [consulté le 13 août 2007], Vol. 9, n° 7-8. <http://www.dlib.org/dlib/july03/young/07young.html>

L'article présente des projets qui utilisent l'OAI-PMH d'une façon originale comme le thésaurus GSAFD, le Digital Library Usage Log ou l'OpenURL Registry. L'article est ardu.

Le format Dublin Core

[10] HILLMANN Diane. Using Dublin Core. DCMI [en ligne], nov. 2005. [consulté le 20 sept. 2007], <http://dublincore.org/documents/usageguide/#basicprinciples>

Cet article propose une première approche de Dublin Core aux utilisateurs qui choisissent ce format.

[11] LAGOZE Carl. Keeping Dublin Core simple: cross-domain discovery or resource description? D-Lib Magazine [en ligne], janv. 2001. [consulté le 20 août 2007]. Vol. 7, n° 1. <http://www.dlib.org/dlib/january01/lagoze/01lagoze.html>

Le format Dublin Core a été créé dans la perspective de réaliser des recherches inter-domaine. Or, la création du Dublin Core Qualifié née du besoin d'adapter le DC à sa communauté de pratiques, va à l'encontre des fondements du DC.

[12] WOODLEY Mary. DCMI Glossary. DCMI [en ligne], nov. 2005. [consulté le 20 sept. 2007], <http://dublincore.org/documents/usageguide/glossary.shtml>

Glossaire en ligne réalisé à partir des documents de la DCMI (recommandations, présentations, articles des conférences Dublin Core)

[13] Dublin Core Metadata Element Set, Version 1.1. DCMI [en ligne], déc. 2006. [consulté le 20 sept. 2007], <http://dublincore.org/documents/dces/>

Ce sont les recommandations de la DCMI (Dublin Core Metadata Initiative) concernant les 15 éléments DC.

Méthodologie

[14] ALEXANDER Jan, TATE, Marsha Ann. Checklist for an information web page. [en ligne]. Widener University, [1996], [consulté le 20 sept. 2007]. <http://www3.widener.edu/Academics/Libraries/Wolfgram_Memorial_Library/Evaluate_Web_Pages/Checklist_for_an_Information_Web_Page/5720/>

Critères d'analyse pour évaluer la fiabilité d'un site à contenu informationnel.

[15] DUPLESSIS Pascal. Inventaire des concepts info-documentaires mobilisés dans les activités de recherche d'informations en ligne [en ligne], Académie de Nantes, déc. 2005, [consulté le 20 septembre 2007]. <http://www.ac-nantes.fr:8080/peda/disc/cdi/reseau/crjrl05/JRL49-4.pdf>

Schémas de modélisation d'une recherche en ligne.

[16] SERRES Alexandre. L'évaluation et ses différents objets. **In** Evaluation de l'information sur internet, Rennes, URFIST de Rennes, 2002, 1er mars 2007, [consulté le 20 sept. 2007] <http://www.uhb.fr/urfist/Supports/evaluationinfo/EvalInfo_evaluation.htm>

Support de formation proposant des grilles d'évaluation de l'information trouvée sur le web. Ces grilles se divisent en quatre catégories : l'auteur, le contenu, la structuration et la mise en forme.

Qualité des métadonnées

[17] HILLMANN Diane, DUSHAY Naomi. Analyzing metadata for effective use and re-use. In DC-2003: 2003 Dublin Core Conference [en ligne]. Sept. 2003, Seattle. [consulté le 13 août 2007] <http://www.cs.cornell.edu/naomi/dc2003/dc2003-dushay-hillmann.pdf>

L'évaluation de larges lots de métadonnées provenant d'un environnement agrégé permet de détecter les problèmes les plus courants. Cet article m'a aidée à structurer ma partie « étude ».

[18] HUTT Arwen, RILEY Jenn. Semantics and syntax of Dublin Core usage in Open Archives Initiative data providers of cultural heritage materials. In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries. 7-11 juin 2005, Denver, Colorado, Etats-Unis. New

York, ACM Press, 2005. p. 262-270

L'article, en rendant compte d'une étude effectuée sur les champs DC Creator*,* Contributor *et* Date *de métadonnées moissonnées en OAI-PMH et appartenant à des organisations du domaine patrimonial, révèle les problèmes que pose un manque de qualité dans les métadonnées pour la recherche.*

[19] PARK Jung-ran. Semantic interoperability and metadata quality: an analysis of metadata item records of digital image collections. Knowledge Organization, 2006, vol. 11, n° 1, p.20-34.

L'article rend compte d'une étude menée sur les métadonnées au format DC et fait ressortir les éléments qui font naître le plus d'ambiguïtés et d'incohérences.

[20] SHREEVES Sarah L. [et al]. Is « quality » metadata « shareable » metadata? The implications of local metadata practice on federated collections. In Proceedings of the Twelfth National Conference of the Association of College and Research Libraries [en ligne]. 7-10 avril 2005, Minneapolis, Etats-Unis. [consulté le 20 août 2007] <http://www.ala.org/ala/acrl/acrlevents/shreeves05.pdf>

Les métadonnées d'un environnement agrégé, pour être de qualité, doivent être partageables ; elles doivent présenter des informations permettant de les comprendre dans leur nouveau contexte. C'est un article essentiel pour comprendre ce qu'est une métadonnée de qualité.

[21] SHREEVES Sarah L. [et al]. Moving towards shareable metadata. First Monday [en ligne], août 2006. [consulté le 17 août 2007], Vol. 11, n° 8. <http://www.firstmonday.org/issues/issue11_8/shreeves/index.html>

Dans un environnement agrégé, le manque de qualité et d'interopérabilité des métadonnées ne permet souvent pas de dépasser des fonctionnalités de recherche très simples. L'article propose une définition de ce que sont des métadonnées partageables.

Bonnes pratiques

Par portail étudié

[22] VAN VEEN Theo, OLDROYD Bill. Search and retrieval in the European Library: a new approach. D-Lib Magazine [en ligne], février 2004. [consulté le 16 août 2007], Vol. 10, n° 2. <http://www.dlib.org/dlib/february04/vanveen/02vanveen.html>

Cet article très technique permet de connaître le fonctionnement général de TEL et le format de description choisi.

[23] Best Practices for Shareable Metadata. [en ligne]. Digital Library Federation, [août 2005], juin 2007 [consulté le 20 sept. 2007]. <http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/ShareableMetadataPublic>

Ce wiki réunit des bonnes pratiques concernant les métadonnées autour de la notion de partageabilité. On y trouve des préconisations au niveau du contenu (conseils sur les formats, l'utilisation de vocabulaires contrôlés) mais aussi au niveau des aspects techniques (schémas XML, encodage). C'est un outil très riche, libre d'accès mais modifiable par un certain profil d'utilisateurs.

[24] The European Library Handbook. The European Library, [s. d.], 27 juin 2007, [consulté le 20 sept. 2007]. <http://www.theeuropeanlibrary.org/handbook/handbook.php>

C'est un guide très complet permettant aux bibliothèques/fournisseurs de données de préparer et traiter leurs métadonnées, de les mettre à disposition dans un entrepôt afin de les rendre moissonnables par TEL. La partie qui nous intéresse plus particulièrement traite des préconisations concernant les métadonnées et l'utilisation de vocabulaires contrôlés et de normes. Cet outil est également très riche et libre d'accès.

[25] Information for data providers. Sheet Music Consortium, [s. d.], 1er juin 2006, [consulté le 20 sept. 2007]. <http://digital.library.ucla.edu/sheetmusic/OAIProject.html>

Guide de préconisations à l'attention des fournisseurs de données, le document propose un tableau reprenant les 15 éléments DC et expliquant pour chacun comment traiter ses métadonnées.

Généralités

[26] COLE Timothy W. [et al]. Now that we've found the « hidden web », what can we do with it? University of Illinois at Urbana-Champaign [en ligne], 2002. [consulté le 10 août 2007], <http://www.archimuse.com/mw2002/papers/cole/cole.html>

Après une présentation de l'utilisation de l'OAI-PMH à l'Université de l'Illinois, l'article développe des conseils relatifs à la granularité, aux métadonnées et aux vocabulaires contrôlés.

[27] DUSHAY Naomi, HILLMANN Diane. NSDL Metadata Primer. [en ligne]. NSDL, 2005, [consulté le 20 sept. 2007]. <http://metamanagement.comm.nsdlib.org/outline.html>

Ces recommandations sur les métadonnées, bien qu'à destination des partenaires du projet de la NSDL, sont tout à fait généralisables.

[28] HAGEDORN Kat [et al]. Enhancing Search and Browse Using Automated Clustering of Subject Metadata. D-Lib Magazine [en ligne], juillet/août 2007. [consulté le 13 août 2007], Vol. 13, n° 7-8. <http://www.dlib.org/dlib/july07/hagedorn/07hagedorn.html>

L'article rend compte de la méthode utilisée par les responsables d'OAIster pour catégoriser a posteriori les métadonnées moissonnées.

[29] HILLMANN Diane. Getting the word out: making digital project metadata available to aggregators. First Monday [en ligne], août 2006. [consulté le 17 août 2007], Vol. 11, n° 8. <http://www.firstmonday.org/issues/issue11_8/hillmann/index.html>

L'OAI-PMH est un moyen pour les bibliothèques de faire connaître leurs ressources. Mais il est nécessaire, pour être visible dans un environnement agrégé, de préparer ses métadonnées au moissonnage.

[30] HILLMANN Diane, DUSHAY Naomi. Improving metadata quality: augmentation and recombination. In DC-2004: International Conference on Dublin Core and Metadata Applications [en ligne]. Oct. 2004, Shangai. [consulté le 13 août 2007] <http://www.cs.cornell.edu/naomi/DC2004/MetadataAugmentation--DC2004.pdf>

L'unité de travail de base dans l'utilisation d'OAI-PMH est l'enregistrement. Or, la NSDL, en tant que fournisseur de services, tente d'améliorer les métadonnées qu'elle moissonne en les traitant au niveau de l'élément ou du champ. Ce document est à destination des fournisseurs de services qui veulent améliorer les métadonnées qu'ils ont récoltées.

[31] LIU Xiaoming [et al]. ARC, an OAI service provider for Digital Library Federation. D-Lib Magazine [en ligne], avril 2001. [consulté le 10 août 2007], Vol. 7, n° 4. <http://www.dlib.org/dlib/april01/liu/04liu.html>

Retour d'expérience sur l'implémentation d'un portail par un fournisseur de services, le document insiste sur l'importance de recevoir des métadonnées normalisées pour mettre en place une recherche efficace.

[32] LIU Xiaoming [et al]. Lessons learned with Arc, an OAI-PMH service provider. Library Trends, printemps 2005, vol. 53, n° 4, p. 590-603.

Suite du retour d'expérience (cf notice précédente)*, l'article propose des bonnes pratiques à l'usage des fournisseurs de services mais également des fournisseurs de données.*

[33] TENNANT Roy. Bitter harvest: problems & suggested solutions for OAI-PMH data & service providers. California Digital Library [en ligne], [2006]. [consulté le 10 août 2007]. <http://www.cdlib.org/inside/projects/harvesting/bitter_harvest.html>

L'auteur identifie des problèmes liés à la mise en place de portails utilisant l'OAI-PMH et propose des bonnes pratiques à l'attention des fournisseurs de données et de services.

[34] TENNANT Roy. Doing data differently. Library Journal.com [en ligne], juin 2005. [consulté le 17 août 2007]. <http://www.libraryjournal.com/article/CA606393.html>

Le texte revient sur l'importance de normaliser et d'enrichir ses métadonnées quand elles sont présentées dans un environnement agrégé.

[35] ZENG Marci Lei, CHAN Lois Mai. Metadata interoperability and standardization – A study of methodology part II: achieving interoperability at the record and repository levels. D-Lib

Magazine [en ligne], juin 2006. [consulté le 13 août 2007], Vol. 12, n° 6. <http://www.dlib.org/dlib/june06/zeng/06zeng.html>

L'article rend compte d'une méthode d'amélioration de l'interopérabilité de métadonnées agrégées au niveau de l'enregistrement et de l'entrepôt. Il développe une méthode permettant la création d'une notice à partir de notices provenant de différents FD.

Glossaire

Toutes les définitions proviennent du document *Collaborer pour de nouveaux services culturels en ligne* de Muriel Foulonneau (notice n° 4) sauf indication contraire.

Agrégateur : Rassemble les métadonnées provenant de plusieurs fournisseurs de données et les rend accessibles dans un entrepôt OAI.

Architecture distribuée ou répartie : Système dans lequel les données nécessaires à une application ou un service sont localisées dans plusieurs emplacements.

Base de données : Ensemble structuré d'éléments d'information, généralement agencés sous forme de tables, dans lesquels les données sont organisées selon certains critères en vue de permettre leur exploitation.

Définition du Grand Dictionnaire terminologique <www.granddictionnaire.com>

Communauté de pratiques : groupe de personnes qui partagent un sujet d'intérêt commun et qui approfondissent leurs connaissances dans ce domaine en interagissant de manière régulière. Elle se caractérise par l'engagement mutuel des membres, leur entreprise commune et un répertoire partagé.

STENUIT Florence. Créer et animer des communautés de pratique : préconisations pour une entreprise de formation et de conseil. 2006. 200 p. Mémoire DESS, INTD. 2006

Entrepôt OAI : Base de métadonnées constituée par un fournisseur de données. Les métadonnées y sont disponibles dans différents formats afin de répondre à différents types de demandes.

Format de description : Ensemble de champs ou d'éléments servant à décrire et à identifier un document.

Fournisseur de données : Détient des contenus et les met à la disposition d'un FS pour la réalisation d'une application. Il crée un entrepôt de métadonnées OAI.

Fournisseur de services : Lance un programme (le moissonneur) pour collecter les métadonnées d'un ou plusieurs FD et les agrège pour créer un service (application).

HTTP (HyperText Transfer Protocol) : Protocole utilisé pour transférer des documents hypertextes ou hypermédias entre un serveur Web et un client Web.

Définition du Grand Dictionnaire terminologique <www.granddictionnaire.com>

Interopérabilité organisationnelle : Les modes d'organisation sont similaires ; les processus de mise à jour des données sont similaires dans divers établissements, les personnes qui assurent la maintenance ont des fonctions ou qualifications comparables.

Interopérabilité sémantique et syntaxique : Les métadonnées sont similaires ou comprennent des liens d'équivalence car elles représentent les mêmes concepts. Leur syntaxe est similaire ou une procédure d'équivalence existe.

Interopérabilité technique : Les systèmes utilisés peuvent communiquer grâce à des protocoles et langages similaires ou pour lesquels il existe une procédure d'équivalence.

Lot : Groupe d'éléments définis dans un entrepôt. Ces ensembles peuvent être divisés en sous-ensembles.

Mapping : L'équivalent français pourrait être "mise en correspondance" ou "mappage". Opération qui consiste à associer les données d'un ensemble aux données d'un autre ensemble (par exemple, équivalence des champs DC avec un autre format).

Définition du Grand Dictionnaire terminologique <www.granddictionnaire.com>

Métadonnées : Ensemble structuré d'informations décrivant une ressource.

Modèle d'entrepôt : un modèle **partageable** est un modèle ouvert d'entrepôts qui visent à proposer des métadonnées à tout moissonneur qui en fait la demande. Il implique une volonté de diffuser ses données auprès du plus grand nombre possible de moissonneurs.

Un modèle **réservé** se rapporte à des entrepôts conçus pour l'utilisation des données par un service spécifique, au sein d'un projet basé sur une communauté. Ils ne sont créés que dans le but d'alimenter ce service.

Moissonneur : Programme lancé par le fournisseur de services pour la collecte de métadonnées auprès d'un ou plusieurs entrepôts OAI.

Normaliseur : Application d'un FS ou d'un agrégateur qui retraite les données agrégées pour les rendre utilisables par le service.

Notice d'autorité : Notice descriptive qui présente la forme officielle d'un nom ou d'un énoncé de sujet utilisé comme clé d'accès, les formes alternatives, les formes rejetées, les formes associées, des notes explicatives, des notes historiques et toute autre information utile.

Définition de l'EBSI (École de bibliothéconomie et des sciences de l'information), Montréal

Point d'accès : Critères de recherche (titre, auteur, date,...). Ils servent de points d'entrée par lesquels l'utilisateur accède à une ressource.

Portail (dans le présent mémoire) : Site web fournissant un point d'entrée unique à des ressources d'information multiples et disséminées. L'accès aux ressources est réalisé à travers une recherche par mots-clés et/ou de type répertoire.

Protocole : Série de questions posées par une machine, auxquelles une autre machine peut renvoyer une série de réponses. Ce langage commun forme un protocole. Un système capable de comprendre ce langage et de l'utiliser est compatible avec le protocole. Ex : HTTP, XML, OAI-PMH.

Schéma (modèle) de métadonnées : Un modèle de données est un ensemble de champs ou éléments qui permettent de décrire une ressource. Il est exprimé sous la forme d'un certain nombre d'éléments et de règles syntaxiques qui peuvent être encodées en SGML (DTD) ou en XML (schéma XML) ou simplement sous la forme d'un texte normatif.

URL : Chaîne de caractères normalisés servant à identifier et à localiser des ressources consultables sur Internet et à y accéder à l'aide d'un navigateur.

Usabilité / utilisabilité : Capacité d'un système à permettre à ses utilisateurs de faire efficacement ce pourquoi ils l'utilisent. Afin que le travail soit fait, le système « utilisable » doit non seulement être facile à utiliser, mais aussi fiable et efficace.

Web 2.0 : Evolution du web vers une plateforme informatique fournissant des applications web aux utilisateurs. Les applications qui illustrent le mieux ce concept sont les blogs et les wikis.

Web invisible (également web profond, web caché) : Partie du web correspondant à l'ensemble des documents qui ne sont pas indexés par les outils de recherche traditionnels.

Définition du Grand Dictionnaire terminologique <www.granddictionnaire.com>

Web sémantique : web intelligent dans lequel les informations, auxquelles on donne une signification bien définie, sont reliées entre elles de façon à ce qu'elles soient comprises par les ordinateurs, dans le but de transformer la masse des pages web en un index hiérarchisé et de permettre de trouver rapidement les informations recherchées.

Définition du Grand Dictionnaire terminologique <www.granddictionnaire.com>

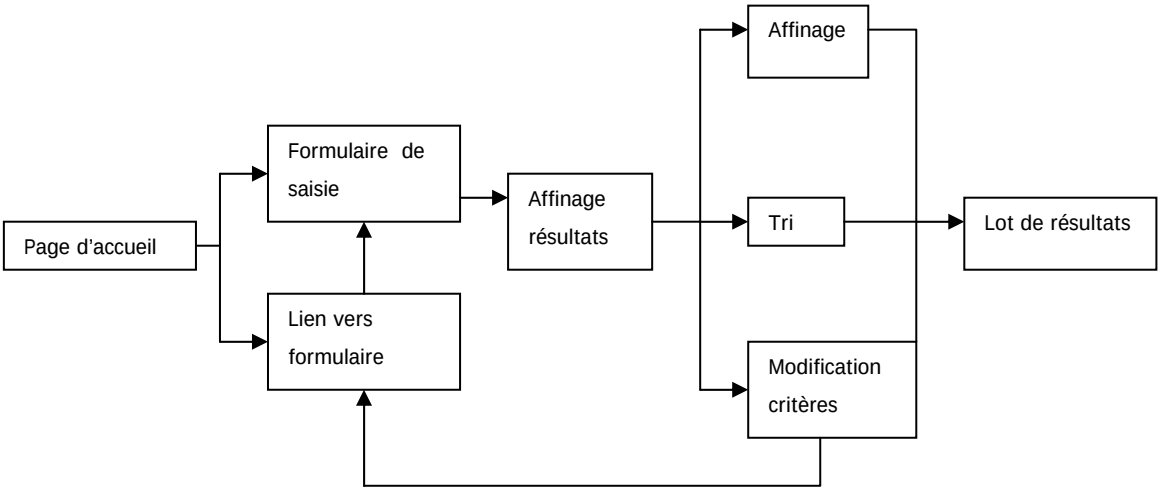
XML (eXtensible Markup Language) : Langage de balisage dérivé du langage SGML, conçu pour faciliter la modification et la validation des programmes qui en découlent, et principalement utilisé pour l'échange d'information entre des systèmes informatiques hétérogènes. C'est un métalangage qui permet de séparer le contenu d'un document de sa présentation et de définir son propre langage pour décrire ce contenu.

Le langage XML a fait l'objet d'une recommandation du consortium W3C. Parmi les nouveaux langages de balisage fondés sur le langage XML, on peut mentionner : RDF, RSS, MathML, XHTML, SVG et cXML.

Définition du Grand Dictionnaire terminologique <www.granddictionnaire.com>

Annexes

Annexe 1 Modélisation d'une recherche en ligne



Etapes d'une recherche

• Recherche simple • Recherche avancée	• Aides à la saisie (index, thésaurus) • Aide à la recherche	• Agencement des résultats • Format (liste, notices allégées, notices complètes) • Classement des résultats	• Présence de délimiteurs • Tri par champ	• Agencement des résultats • Format (liste, notices allégées, notices complètes) • Classement des résultats

Fonctionnalités associées

Annexe 2 Grille d'évaluation de sites web

Présentation générale du portail		
Fiche d'identité	Nom	
	URL	
	Producteur	
	Secteur d'activité	
	Nature (universitaire, institutionnel, commercial,...)	
	Statut (public, privé)	
	Public visé	
	Objectifs	
	Fournisseurs de données	
	Mise en ligne	
	Date de dernière mise à jour	
Autorité	Responsabilité du contenu clairement identifiée ?	
	Sérieux, fiabilité de l'organisation ?	
	Lien vers une page explicitant qui est l'auteur ?	
	Adresse auteur ?	
	Lien vers FD ?	

Annexe 3 Grille d'évaluation d'une recherche en ligne

Fonctionnalités de recherche		
Formulaire de recherche simple	Critères ?	
	Aides à la saisie ?	
	Aides à la recherche ?	
Formulaire de recherche avancée	Critères ?	
	Variété des critères ?	
	Critères permettant une recherche fine ?	
	Opérateurs booléens ?	
	Possibilité de rechercher par collection/FD ?	
	Possibilité de limiter la recherche aux seuls contenus numériques ?	
	Aides à la saisie ?	
	Aides à la recherche ?	
Affichage des résultats	Présentation (densité, clarté, homogénéité)	
	Bruit ? Silence ?	
	Classement	

	Format d'affichage	
	Granularité. Différents niveaux de documents liés entre eux ?	
	Possibilité de tri ?	
	Possibilité d'affiner par délimiteurs ?	
	Distinction claire entre notices bibliographiques simples et notices avec consultation du document numérique ?	
	Liens externes clairement identifiés ?	
	Lien au document numérique clairement indiqué ?	
	Accès au document numérique aisé ?	
	Frustration de l'utilisateur à l'issue de la recherche ? (accès à notices seules, pas distinction notices/ressources,...)	
	Format de métadonnées	

Annexe 4 Grilles d'évaluation renseignées

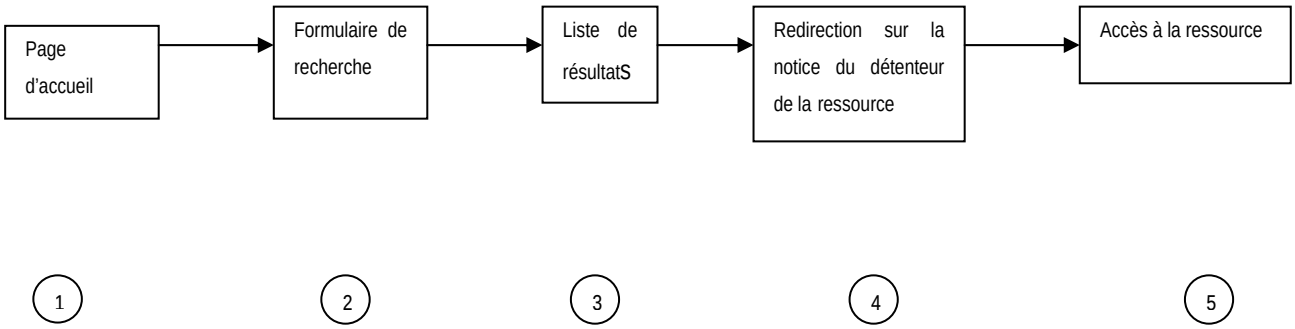
Sheet Music Consortium

PRESENTATION GENERALE DU PORTAIL		
Fiche d'identité	Nom du portail	Sheet Music Consortium
	URL	http://digital.library.ucla.edu/sheetmusic/
	Producteur	UCLA (+ John Hopkins Uni, Library of Congress, Indiana Uni)
	Secteur d'activité	Musique
	Nature du portail (universitaire, institutionnel, commercial, associatif,…)	Universitaire
	Statut du site (public, privé)	Public
	Public visé	Chercheurs, musiciens
	Objectifs du site	Créer une collection numérique de musique en feuilles
	Fournisseurs de données	7 bibliothèques américaines
	Mise en ligne	Sept. 2003
	Dernière mise à jour	1er juin 2006
Autorité	Responsabilité du contenu clairement identifiée ?	Oui
	Sérieux, fiabilité de l'organisation	Oui
	Lien vers une page explicitant qui est l'auteur ?	Oui
	Adresse de l'auteur présente ?	Non
	Liens vers les fournisseurs de données ?	Oui, avec nombre de documents mis à disposition par fournisseur de données

FONCTION RECHERCHE		
Formulaire de recherche simple	Critères de recherche ?	Recherche par mots-clés avec possibilité de limiter aux seuls contenus numériques
	Aide à la saisie (index, thésaurus) ?	Non
	Aide à la recherche ?	Oui, affichée sur la page
Formulaire de recherche avancée	Critères de recherche ?	• 4 zones de saisie avec comme critères Mots-clés, Compositeur/Parolier, Titre, Sujet, Editeur • Date de publication, Collection (liste déroulante) • Tri par titre ou date, Nombre de résultats par page, Contenu numérique seulement
	Variété et nombre des critères de recherche ?	OK
	Critères permettant une recherche fine et précise ?	Oui
	Opérateurs booléens ?	Oui
	Possibilité de rechercher par collection/par fournisseur de données ?	Oui
	Possibilité de limiter la recherche aux seuls contenus numériques ?	Oui
	Aide à la saisie (index, thésaurus) ?	Menu déroulant pour les collections
Affichage des résultats	Présentation des résultats (densité, clarté, homogénéité,…)	OK, le plus : choix du nombre de résultats par page
	Classement des résultats	Choix
	Format d'affichage (liste, notice allégée, notice complète) ?	Liste homogène avec critères suivants : Titre, Créateur, Editeur, Collection
	Granularité ? Différents niveaux de document liés entre eux ?	Niveau de granularité fixe (partition). Apparition d'un champ Relation indiquant la collection d'appartenance de la ressource
	Possibilités de tri ? Si oui, lesquels ?	Oui, cf supra
	Possibilité d'affiner la	Non

	recherche sans repasser par le formulaire de recherche ? Si oui, quels délimiteurs ?	
	Distinction claire entre les notices bibliographiques simples et les notices avec consultation du document numérique ?	Oui car possibilité de ne rechercher que dans les collections numériques et mention [access online] indiquant le lien à la ressource numérique
	Liens externes clairement identifiés ?	Oui
	Lien au document numérique clairement indiqué ?	Oui
	Accès au document numérique aisé ?	Oui
	Frustration de l'utilisateur à l'issue de la recherche ? (accès à notices seules, pas distinction notices/ressources,…)	Non

Modélisation du chemin vers le document numérique



Europeana

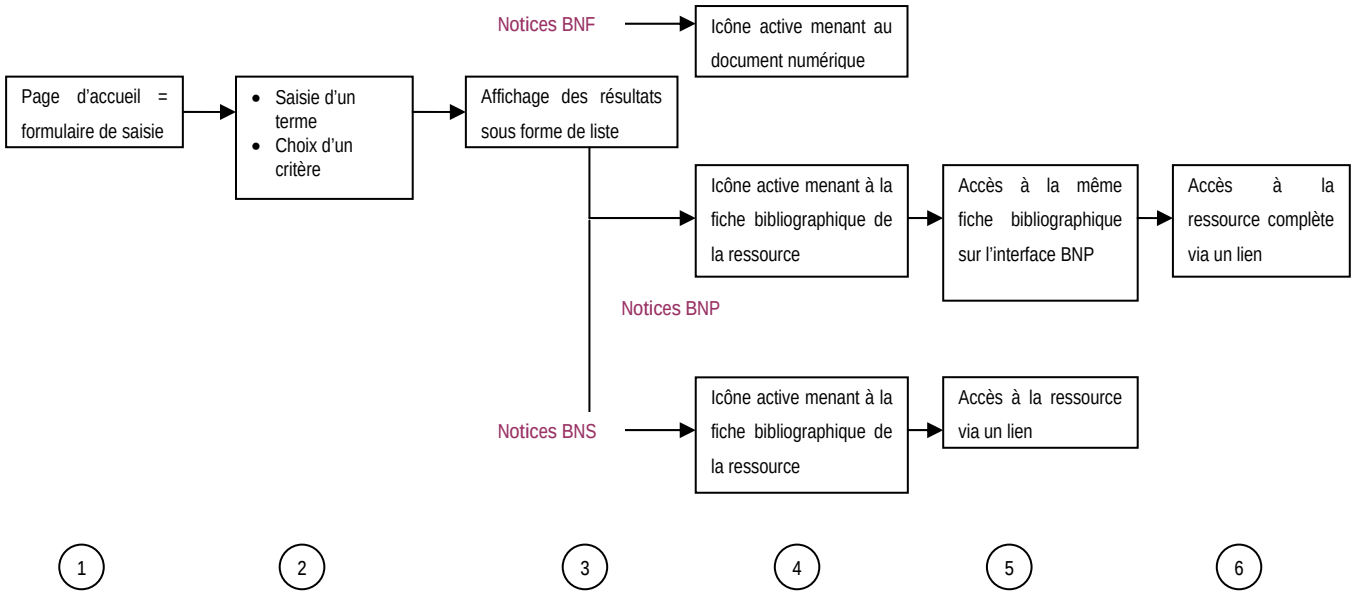
PRESENTATION GENERALE DU PORTAIL		
Fiche d'identité	Nom du portail	Europeana
	URL	http://www.europeana.eu/
	Producteur	BNF
	Secteur d'activité	Bibliothèque
	Nature du portail (universitaire, institutionnel, commercial, associatif,…)	Institutionnel
	Statut du site (public, privé)	Public
	Public visé	Public large et diversifié
	Objectifs du site	Prototype de bibliothèque en ligne (préfigure la Bibliothèque Numérique Européenne) Expérimente de nouvelles fonctionnalités Possibilité pour les utilisateurs de poster des commentaires
	Fournisseurs de données	BNF, Bibliothèque Nationale Széchényi de Hongrie et Bibliothèque nationale du Portugal
	Mise en ligne	Mars 2007
	Dernière mise à jour	?
	Responsabilité du contenu clairement identifiée ?	Oui, texte sur la gauche de l'écran (lu en premier)
Autorité	Sérieux, fiabilité de l'organisation	Oui
	Lien vers une page explicitant qui est l'auteur ?	Oui (lien « BNF » + lien « A propos » avec information très complète sur le projet)
	Liens vers les fournisseurs de données ?	Oui, en page d'accueil sur l'encart de gauche

FONCTION RECHERCHE		
Formulaire de recherche	Formulaire de recherche présent ? Si oui, où dans la page ?	Recherche par mots-clés : • Champ avec une flèche très simple • Situé tout en haut et au milieu de l'écran • Pas de libellé Recherche par points d'accès : • Deux blocs de sélection de points d'accès avec menu déroulant : o « Choisissez un critère » : Epoque de publication (siècles), Langue, Provenance (bibliothèques) o « Choisissez un thème » : Généralités, Philosophie et Psychologie, Religion, Economie et Société, Langue, Sciences, Techniques, Arts et loisirs, Littérature, Histoire et Géographie
	Si non, liens vers la recherche visibles ?	/
	Recherche simple et avancée ?	Recherche simple seulement, pas de liens à un formulaire de recherche avancée
	Critères de recherche ?	Oui, cf supra « Formulaire de recherche présent ? »
	Aide à la saisie (index, thésaurus) ?	Non
	Aide à la recherche ?	Oui, « Aide en ligne », en haut à droite sous forme d'un point d'interrogation. Ouvre une page d'aide. On apprend qu'il y a 3 types de recherche simple (ce qu'on ne comprend pas d'emblée !) • Plein texte • Par critères • Par thèmes
	Variété et nombre des critères de recherche ?	OK mais non combinables dès la page d'accueil…
	Critères permettant une recherche fine et précise ?	Oui, grâce aux possibilités d'affinage mais seulement après avoir effectué une première recherche très large
	Opérateurs booléens ?	Non
	Possibilité de rechercher par collection/par fournisseur de données ?	Oui, par bibliothèque mais pas par collection (d'ailleurs, y a-t-il par bibliothèque plusieurs collections ?)

	Possibilité de limiter la recherche aux seuls contenus numériques ?	Tous les contenus numériques décrits sont accessibles
	Aide à la saisie (index, thésaurus) ?	Index d'auteurs à l'affichage des résultats à gauche de l'écran
	Aide à la recherche ?	Non
	Présentation des résultats (densité, clarté, homogénéité,…)	Densité OK (10 résultats en liste par page) Clarté : moyenne (tout est en gras donc rien ne ressort)
	Classement des résultats	Pas indiqué. Il faut le déduire…
Affichage	Format d'affichage (liste, notice allégée, notice complète) ?	Affichage premier : format liste, pas homogène (même au sein de BNF) ➔ Titre / Auteur parfois / Date ➔ Champs Titre et Auteur séparés par virgule ou par slash ➔ Une seule constante : date indiquée à chaque fois et séparée par un tiret Possibilité d'afficher les notices complètes mais une par une (même icône que « A propos »…)
	Granularité ? Différents niveaux de document liés entre eux ?	Non. Un niveau de granularité, le livre numérisé
	Possibilités de tri ? Si oui, lesquels ?	Oui, affinage
	Possibilité d'affiner la recherche sans repasser par le formulaire de recherche ? Si oui, quels délimiteurs ?	Oui, dans une palette déplaçable située à gauche. (recherche par facettes) En fait, c'est plutôt une fonction « Tri » par autres critères (Provenance, Auteur, Langue) Mais pas de croisement avec les thèmes par exemple. En revanche, quand recherche par « Thème », croisement avec « Critères » possible
	Possibilité de modifier les critères de la recherche (bouton renvoyant au	Pas nécessaire vu que la recherche est simple et que l'entrée se fait par un seul critère. Retour à l'accueil suffit.

	formulaire précédent) ?	
	Distinction claire entre les notices bibliographiques simples et les notices avec consultation du document numérique ?	Tous les doc. numériques sont consultables (sur interface Europeana pour la BNF, sur site d'origine pour les deux autres)
	Liens externes clairement identifiés ?	Non. Quand consultation de la ressource, pour BNF on reste sur interface d'Europeana, pour les autres, on sort du site (sans ouverture d'une nouvelle fenêtre, ce qui pose pb en terme de navigation) Cela est explicité dans « A propos » mais ça ne suffit pas Par ailleurs, problématique au niveau des droits…
	Lien au document numérique clairement indiqué ?	Oui
	Accès au document numérique aisé ?	Oui
	Frustration de l'utilisateur à l'issue de la recherche ? (accès à notices seules, pas distinction notices/ressources,…)	Non, que des documents numériques

Modélisation du chemin vers le document numérique



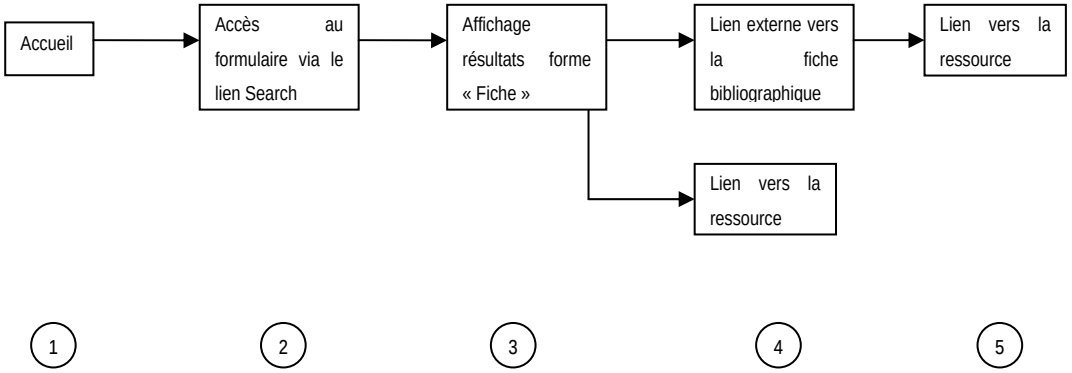
OAIster

PRESENTATION GENERALE DU PORTAIL		
Fiche d'identité	Nom du portail	OAIster
	URL	http://www.oaister.org/
	Producteur	University of Michigan Partenaire : University of Illinois Urbana Champaign (UIUC) Mellon Fondation (financement d'origine)
	Secteur d'activité	Culture (portail encyclopédique)
	Nature du portail (universitaire, institutionnel, commercial, associatif,…)	Universitaire
	Statut du site (public, privé)	Public
	Public visé	Universitaire
	Objectifs du site	Donner accès à des notices et des ressources de « type universitaire » (*academic resources*)
	Fournisseurs de données	852 fournisseurs
	Mise en ligne	Juin 2002
	Dernière mise à jour	11 juillet 2007
Autorité	Responsabilité du contenu clairement identifiée ?	Oui
	Sérieux, fiabilité de l'organisation	Oui
	Lien vers une page explicitant qui est l'auteur ?	Oui
	Liens vers les FD ?	Oui (moteur de recherche pour trouver un fournisseur de données)

FONCTION RECHERCHE		
Formulaire de recherche	Critères de recherche ?	Trois blocs de critères combinables : • Recherche dans les champs = 3 zones de saisie avec les critères combinables (opérateurs booléens) suivants: o Title o Author/Creator o Subject o Language • Délimiteurs “type de ressources” (texte, image, …) • Critères de tri pour la sortie des résultats (titre, auteur, date, Hit frequency et Weight frequency)
	Variété et nombre des critères de recherche ?	OK
	Critères permettant une recherche fine et précise ?	Oui
	Opérateurs booléens ?	Oui
	Possibilité de rechercher par collection/par fournisseur de données ?	Non, mais affinage par FD sur l’interface des résultats
	Possibilité de limiter la recherche aux seuls contenus numériques ?	Tous les contenus sont numériques
	Aide à la saisie (index, thésaurus) ?	Non
	Aide à la recherche ?	Oui
Affichage des résultats	Présentation des résultats (densité, clarté, homogénéité,…)	Peu claire car pas d’affichage liste et beaucoup de bruit
	Classement des résultats	Possibilité de choisir dès la recherche
	Format d’affichage (liste, notice allégée,	Notices complètes

	notice complète) ?	
	Granularité ? Différents niveaux de document liés entre eux ?	Oui, mais doc pas liés. En revanche, permet de replacer le doc dans son contexte (Note : IsPartOf)
	Possibilités de tri ? Si oui, lesquels ?	Oui, cf Recherche
	Distinction claire entre les notices bibliographiques simples et les notices avec consultation du document numérique ?	Contenu numérique seul
	Liens externes clairement identifiés ?	Oui (adresse URL + ouverture d'une nouvelle fenêtre)
	Lien au document numérique clairement indiqué ?	Oui, par un lien URL
	Accès au document numérique aisé ?	Oui
	Frustration de l'utilisateur à l'issue de la recherche ? (accès à notices seules, pas distinction notices/ressources,…)	Oui, si droits réservés… Il existe un champ Rights mais souvent pas assez clair…

Modélisation du chemin vers le document numérique



The European Library

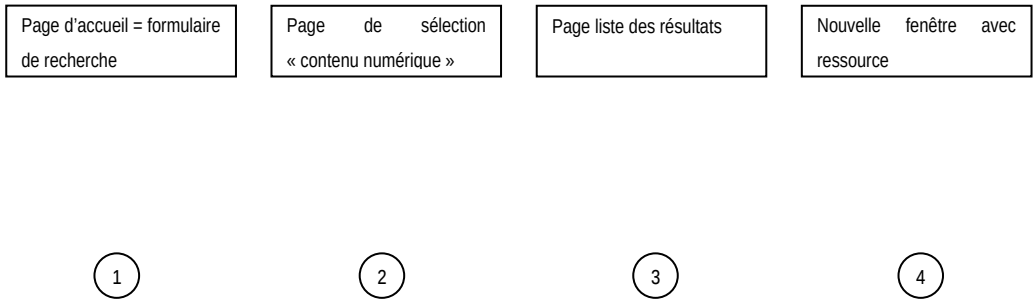
PRESENTATION GENERALE DU PORTAIL		
Fiche d'identité	Nom du portail	The European Library (TEL)
	URL	www.theeuropeanlibrary.org
	Producteur	CENL (Conférence Européenne des directeurs de Bibliothèques Nationales)
	Secteur d'activité	Bibliothèque
	Nature du portail (universitaire, institutionnel, commercial, associatif,…)	Associatif
	Statut du site (public, privé)	Fondation
	Public visé	« Tout citoyen ». Public large et diversifié
	Objectifs du site	• Accès simplifié à des références bibliographiques et à des ressources numériques et numérisées • Accroissement de la visibilité des bibliothèques nationales européennes
	Fournisseurs de données	Collections numériques des bibliothèques nationales de 24 pays européens
	Mise en ligne	Mars 2005
	Dernière mise à jour	/
Autorité	Responsabilité du contenu clairement identifiée ?	Oui
	Sérieux, fiabilité de l'organisation	Oui
	Lien vers une page explicitant qui est l'auteur ?	Oui, avec détails et liens sur auteur et projet
	Liens vers les fournisseurs de données ?	Oui

FONCTION RECHERCHE		
Formulaire de recherche simple	Critères de recherche ?	• Recherche plein texte avec possibilité de sélectionner un type de documents • Cette sélection pas homogène : on y trouve des types de documents comme *cartes*, *atlas*, *photos* ou *manuscrits* et des thèmes comme *religion*, *musique*, *littérature jeunesse*… • Pas de correspondance entre les langues… • Utilisation d'un clavier virtuel pour les caractères spéciaux • A droite de l'écran, possibilité de chercher dans les collections par pays, de feuilleter les collections par sujet et de feuilleter la description des collections, ce qui ouvre une page avec un formulaire composé de 5 points d'accès (mot-clé, type, langue, couverture spatiale et couverture temporelle). Après sélection de collections, sauvegarde du critère et retour en page d'accueil recherche
	Aide à la saisie (index, thésaurus) ?	Non
	Aide à la recherche ?	Oui
Formulaire de recherche avancée	Critères de recherche ?	En cliquant sur un lien « Recherche avancée », un bloc de critères se déroule (il disparaît en recliquant sur « Recherche avancée ». Critères de recherche : • Titre • Auteur • Sujet • Type • Langue • ISBN • ISSN
	Variété et nombre des critères de recherche ?	Trop de critères, trop d'écrans différents. Difficile de comprendre l'organisation et les possibilités de recherche
	Critères permettant une recherche fine et précise ?	Nombreux critères mais la recherche produit du bruit
	Opérateurs booléens ?	Oui, en recherche avancée et en « Recherche dans la description des collections »
	Possibilité de rechercher par collection/par fournisseur de données ?	Oui, mais ouverture d'une nouvelle page avec sauvegarde des critères. Peu pratique
	Possibilité de limiter la recherche aux seuls contenus	Oui mais difficile de sélectionner ce critère (onglet « Collections », sélection de « Matériaux numérisés »)

	numériques ?	
	Aide à la saisie (index, thésaurus) ?	Non
	Aide à la recherche ?	Oui
Affichage des résultats	Présentation des résultats (densité, clarté, homogénéité,…)	• Assez aéré • Homogénéité dans l'affichage des champs en liste (Titre en lien actif qui permet d'afficher la notice complète, auteur en dessous puis type de ressources, ie image, photo, livre, périodique, langue parfois) • Quand une image est consultable, apparaît à droite sous forme de vignette. Consultable en cliquant sur « Voir l'objet » Mais on ne sait pas que l'on sort du site…
	Classement des résultats	Par pays puis par collection
	Format d'affichage (liste, notice allégée, notice complète) ?	Affichage liste assez homogène et affichage complet très variable selon le FD
	Granularité ? Différents niveaux de document liés entre eux ?	Non, granularité variable, sans indication aucune
	Possibilités de tri ? Si oui, lesquels ?	Non
	Possibilité d'affiner la recherche sans repasser par le formulaire de recherche ? Si oui, quels délimiteurs ?	Oui, possibilité d'exclure un terme des résultats, possibilité de rechercher un terme dans les résultats
	Possibilité de modifier les critères de la recherche (bouton renvoyant au formulaire précédent) ?	Oui, en cliquant sur « Recherche avancée » ou sur « Modifier la liste des collections »
	Distinction claire entre les notices bibliographiques simples et les notices avec consultation du document numérique?	Les doc numérisés sortent en premier (liste à gauche) mais pas clair ! Pour les photos, il y a une vignette à droite de la notice. Mais l'intitulé « Voir l'objet » est ambigu car il est présent même quand le lien n'est que vers une notice bibliographique...
	Liens externes clairement identifiés ?	Non
	Lien au document numérique clairement indiqué ?	Non

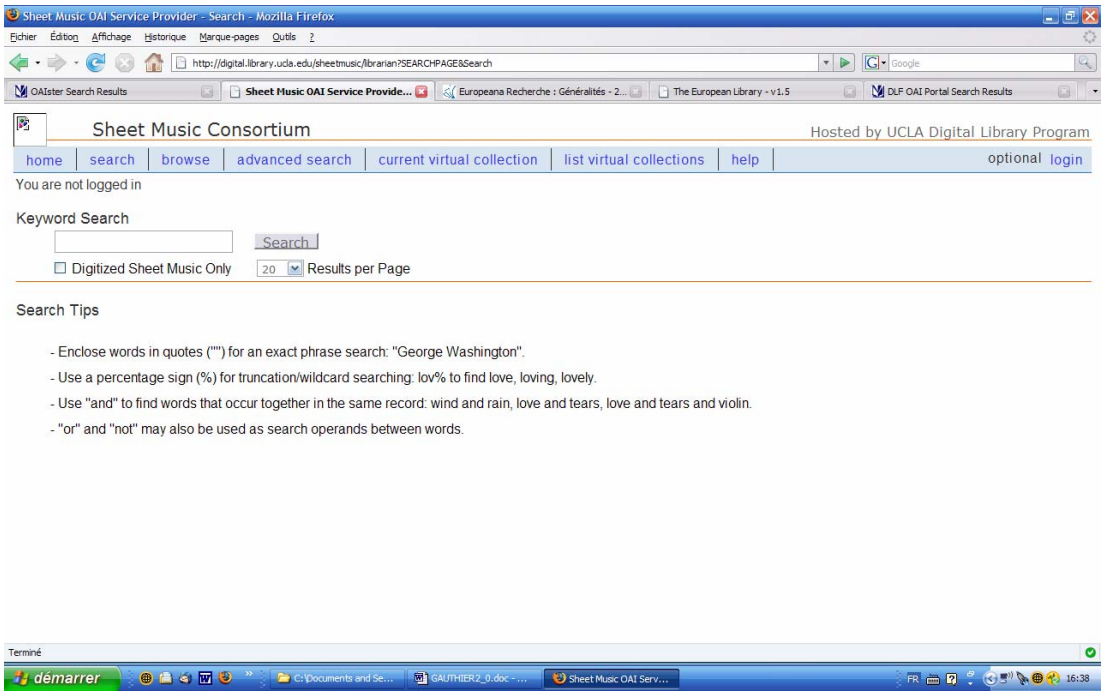
	Accès au document numérique aisé ?	Non
	Frustration de l'utilisateur à l'issue de la recherche ? (accès à notices seules, pas distinction notices/ressources,...)	Oui

Modélisation du chemin vers le document numérique



Formulaires de recherche simple

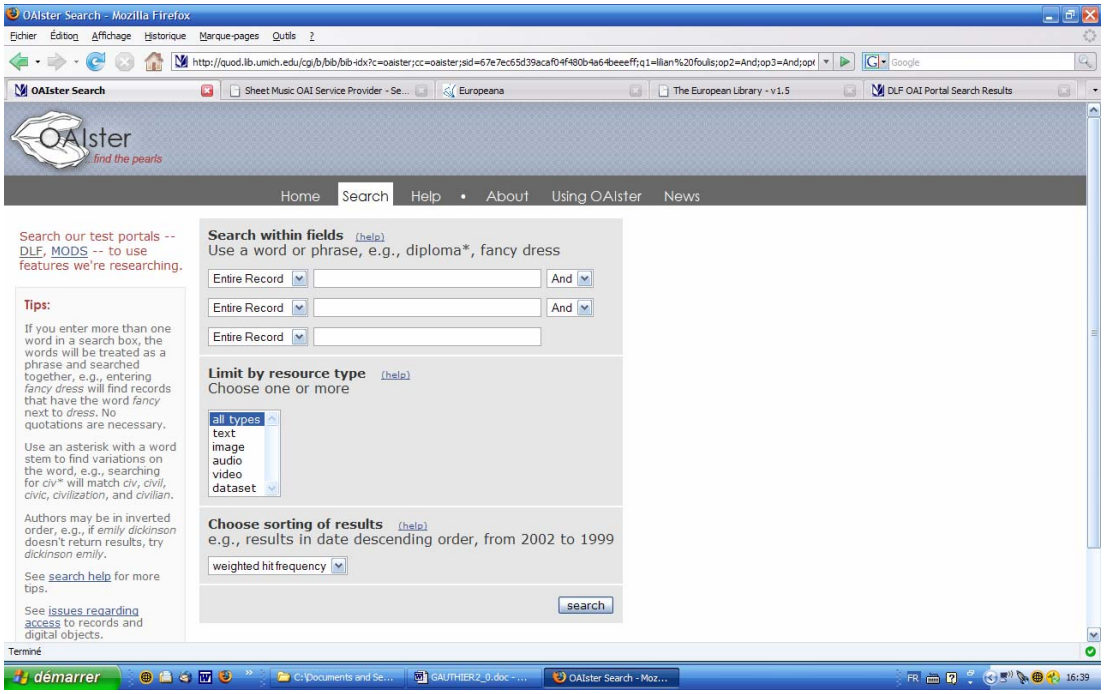
Sheet Music Consortium



Europeana

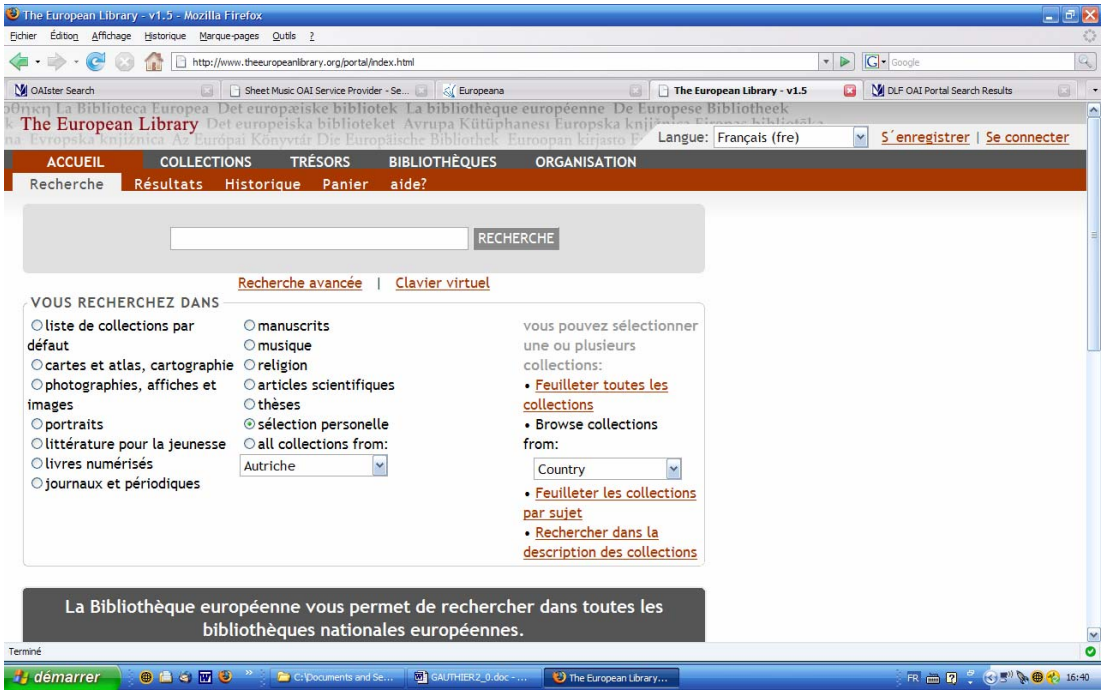
Note : il n'y a qu'un formulaire de recherche sur Europeana

OAIster



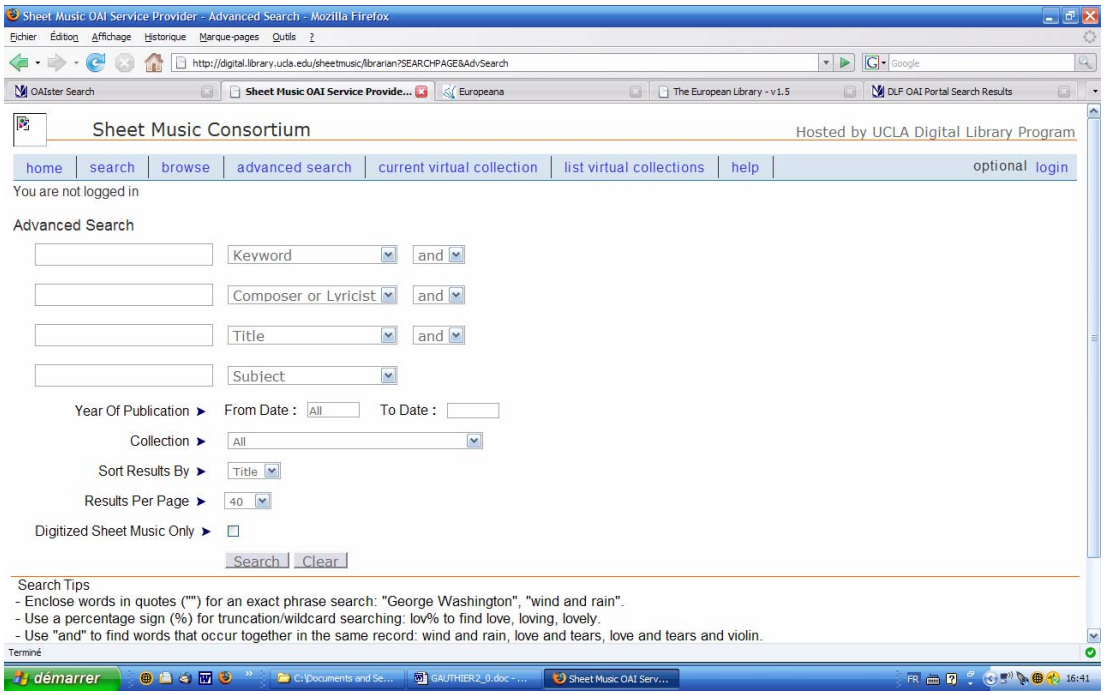
Note : il n'y a qu'un formulaire de recherche sur OAIster

The European Library



Formulaires de recherche avancée

Sheet Music Consortium



The European Library