La recherche fédérée des portails patrimoniaux : quelles solutions documentaires ? : l'exemple du Musée des Civilisations de l'Europe et de la Méditerranée (MuCEM)

Aurélia Giusti

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

Aurélia Giusti

le 6 novembre 2009

La recherche fédérée des portails patrimoniaux : quelles solutions documentaires ?

L'exemple du Musée des Civilisations de l'Europe et de la Méditerranée (MuCEM)

Jury
Jean-Pierre Dalbéra
Ghislaine Chartron

Promotion XXXIX

Remerciements

Je remercie Ghislaine Chartron pour avoir accepté d'encadrer mon travail, pour sa disponibilité, son écoute et ses conseils précieux.

J'exprime envers Jean-Pierre Dalbéra toute ma gratitude pour m'avoir donné la possibilité d'intégrer l'équipe multimédia du MuCEM et la possibilité d'expérimenter de nouvelles technologies. J'ai découvert le métier de documentaliste sous un jour nouveau.

Je remercie Filippo Vancini, Yannick Vernet, Mohan Danabalou, Denis Chevallier, Marina Zveguinzoff pour le temps qu'ils m'ont consacré. Partager mes questionnements avec eux m'a permis d'avancer tout au long de mon stage et de la rédaction de ce mémoire.

Je remercie vivement l'équipe du MuCEM à Paris et à Marseille pour leur accueil chaleureux durant mon stage : Françoise Bekus, Michel Colardelle, Geneviève Deblock, Isabelle Gui, Marie-Barbara Le Gonidec, Emilie Girard, Patrick Dubois, José Albertini, Jacqueline Christophe, Catherine Homo-Lechner, Frederica Tamarozzi, Florence Pizzorni, Marc Touché.

Je remercie Lise Rosat, Emilie Frances pour leur aide concernant le web sémantique et le site Masculin-Féminin.

Je remercie toutes les personnes que j'ai interrogées, qui ont pris le temps de répondre à mes questions et de me recevoir.

Merci à ma famille et mes amis pour leur aide et leur soutien.

Notice

Description bibliograhique :

GIUSTI Aurélia. La recherche fédérée des portails patrimoniaux : quelles solutions documentaires ? L'exemple du MuCEM. Mémoire en vue d'obtenir le diplôme Titre I « chef de projet en ingénierie documentaire », Paris, INTD, 2009, 110 p.

Résumé :

Dans le monde culturel, les bases de données se comptent maintenant par milliers et l'offre d'information est devenue difficile à cerner pour des non-professionnels. Comment faciliter la consultation de ressources hétérogènes aussi vastes et complexes ? La mise en place d'un moteur de recherche fédérée sur un ensemble plus ou moins élevé de bases de données est une réponse à ces difficultés d'accès.

Interroger en une seule requête des ressources hétérogènes demande l'utilisation de protocoles d'interrogation et de métadonnées spécifiques. Si l'interopérabilité technique trouve une solution, qu'en est-il de l'interopérabilité sémantique ? Ce mémoire a pour objectif de présenter les outils utilisés dans la recherche fédérée puis d'analyser les solutions documentaires choisies par différentes institutions culturelles pour répondre à la question de l'harmonisation des mots-clés. La dernière partie fournira des préconisations, à partir d'un cas concret, le futur portail documentaire du MuCEM. Pour finir, des pistes de réflexion seront amorcées concernant les technologies du web sémantique et les possibilités de mettre en correspondance différentes terminologies.

Descripteurs :

Patrimoine numérique culturel ; recherche fédérée ; portail ; interopérabilité ; métadonnée ; langage documentaire ; TAL ; web sémantique

Table des matières

Introduction

La recherche fédérée est la possibilité d'interroger en une seule requête des sources d'information hétérogènes. Les institutions patrimoniales mettent en place cette fonctionnalité dans leur portail documentaire pour diffuser et rendre accessibles leurs collections. En effet, elle permet de faire face à l'augmentation de la masse de documents en ligne et d'accéder aux contenus numériques rapidement et simplement. L'utilisateur n'a pas à consulter les bases une par une, à connaître l'interface et le langage d'indexation de chacune d'elles.

Depuis le milieu de la décennie 70, le ministère de la Culture et de la Communication et ses établissements publics sous tutelle mènent des politiques d'informatisation des inventaires et catalogues. Au début de la décennie 90, la numérisation des fonds documentaires et des documents primaires a commencé dans les musées, archives, bibliothèques, centres de ressources sur le patrimoine mobilier et architectural ou sur le patrimoine photographique, sonore, audiovisuel et cinématographique, etc. Elle se poursuit aujourd'hui. Chaque institution ou service a mis en oeuvre ses propres systèmes informatiques, adaptés aux différents types de fonds patrimoniaux, mais souvent incompatibles entre eux. Dans le monde culturel, les bases de données se comptent maintenant par milliers et l'offre d'information est devenue difficile à cerner pour des non-professionnels. Au-delà de la question de la valorisation et de la diffusion des collections numériques patrimoniales, se pose la question de l'accessibilité de l'ensemble de ces données numériques par le public. Comment faciliter la consultation de ressources hétérogènes aussi vastes et complexes ?

La mise en place d'un moteur de recherche fédérée sur un ensemble plus ou moins élevé de bases de données est une réponse à ces difficultés d'accès.

Pour y parvenir, il faut répondre à un certain nombre de questions : quels outils mettre en place pour rendre ces bases interopérables ? Quel protocole d'interrogation, quel format documentaire et quelles métadonnées vont permettre de faire communiquer les bases entre elles ?

Dans la recherche fédérée, l'usager n'a pas à se préoccuper du langage documentaire qui indexe les documents de chaque base. Mais, lorsqu'il lance une requête avec, par exemple, le mot-clé « mariage », il n'est pas certain que ce terme corresponde à un descripteur d'un thésaurus ou d'une terminologie d'une autre application. Les documents peuvent être indexés avec un autre descripteur comme « noces » puisqu'ils appartiennent à des bases de données différentes utilisant leur propre langage contrôlé. La recherche d'information ne sera pas performante. Lorsqu'il n'y a pas de thésaurus transversal, la recherche fédérée se heurte à l'ambiguïté du langage : un mot peut recouvrir plusieurs réalités, un document peut

être indexé selon différents vocabulaires. Comment les institutions patrimoniales ont-elles fait correspondre les vocabulaires d'indexation propres à chaque base de données ?

Après une première partie définissant la recherche fédérée et ses protocoles d'interrogation, une étude de cas permettra d'observer comment musée, bibliothèque ou portail patrimonial ont réglé la question de l'interopérabilité technique et sémantique.

En 2013, le MuCEM, Musée des civilisations de l'Europe et de la Méditerranée, ouvrira ses portes à Marseille. Ce musée national aura son centre de ressources et proposera à l'utilisateur un outil de recherche fédérée et croisée qui lui permette d'interroger simultanément plusieurs catalogues et inventaires, sur un sujet, une région, un auteur ou un type d'objet et ses documents associés. Son projet est de mettre en place un système d'information cohérent. Héritant de ressources hétérogènes (huit bases de données différentes), comment arriver à faire un système d'information grand public qui permette de retrouver des documents quelque soit le support avec « le bon mot-clé »?

Dans le cadre d'une campagne de collecte de documents (objet, photos, imprimés) intitulé « Mariage », l'équipe scientifique du MuCEM a constitué un corpus sur la construction du genre et les rites de passage qui l'accompagnent. Ce corpus a servi de matériau à un site internet *Masculin-Féminin : la construction du genre* mais aussi à l'expérimentation de différents outils : une bibliothèque numérique au format Dublin Core et des outils de traitement automatique de la langue (TAL). A travers ces tests, s'est posé un certain nombre de questions : comment harmoniser les différents langages contrôlés (thésaurus, index, liste de vedette-matière) ? Le fait de mettre en ligne des recherches jusque là destinées à un public spécialisé oblige-t-il à changer le mode de gestion des collections ?

L'élaboration d'une terminologie avec des outils automatiques de traitement linguistique s'est faite dans le cadre du projet DAFOE, avec la société Mondeca, spécialisée dans le web sémantique, la gestion de thésaurus, taxonomies, terminologies et bases de connaissances multilingues. La plateforme technique DAFOE est un ensemble d'outils comprenant un éditeur d'ontologies. Le futur dira si les technologies du web sémantique ont la capacité de résoudre les problèmes propres des bibliothèques numériques et collections patrimoniales en ligne. Des progrès ont été réalisés en ce sens mais les outils opérationnels restent encore expérimentaux dans les domaines du patrimoine.

La troisième partie sera consacré aux préconisations pour le MuCEM, au projet en cours de création d'une terminologie avec Mondeca, au web sémantique et esquissera une réponse concernant l'harmonisation des langages documentaires.

Première partie
L'interrogation par mot-clé dans la recherche fédérée

1 La recherche fédérée

1.1 Définitions :

1.1.1 Interroger des sources hétérogènes

On peut définir la recherche fédérée comme l'interrogation simultanée, à l'aide d'un formulaire, de ressources numériques hétérogènes.

Par *hétérogène*, on peut comprendre que les ressources interrogées peuvent être de nature différente. Il peut s'agir, d'une part, de ressources primaires : par exemple des travaux universitaires en texte intégral ou des articles de périodiques. Ces ressources sont dites non structurées. Il peut s'agir, d'autre part, de ressources secondaires : métadonnées de catalogues de bibliothèque, notices de bases de données. Ces ressources sont dites structurées.

On peut aussi entendre par *hétérogène* : ressources structurées dans des formats documentaires différents : en MARC pour les catalogues de bibliothèque, en EAD pour les archives…etc.

« La recherche fédérée est avec la gestion de contenu, l'un des deux piliers des portails documentaires » [8, Maisonneuve]. Mais avant d'entrer dans les détails de cette fonctionnalité, il convient de situer la recherche fédérée dans son contexte et, pour cela, de revenir sur quelques notions connexes.

1.1.2 Portail

Le portail est un site web qui offre un point d'accès unique à de multiples services et ressources documentaires. Plusieurs types de site répondent à cette définition. On peut donc introduire d'autres considérations comme le public visé et la raison d'être du portail [8, Maisonneuve].

En ce qui concerne l'information, les portails intègrent des informations internes et des informations externes, ainsi que des informations structurées (bases de données, documents) et des informations non structurées (courrier électronique, messages de forums, notes, etc.).

Les types de portails existants sont nombreux. Une première typologie est la distinction entre portail internet, destiné au public général et accessible par le réseau internet, tel myYahoo !, et le portail d'entreprise, placé sur un intranet ou un extranet et réservé au personnel d'une entreprise. Une autre typologie, qui se place également d'un point de vue de l'usager, distingue le portail horizontal qui a une vocation généraliste du portail vertical, spécialisé autour d'un sujet. Dans le monde du web, les portails horizontaux sont ceux de grands sites comme Yahoo ! et les portails verticaux sont ceux de communautés virtuelles. Au sein d'un organisme, le portail horizontal concerne tout le personnel, alors que le portail vertical ne s'adresse qu'à une partie [10, Stiller].

1.1.3 Portail documentaire

Le portail documentaire reprend certains aspects des portails énumérés plus haut. Il s'agit d'un site web ayant une fonction centralisatrice. Mais ce qui différencie le portail documentaire est qu'il met l'accent sur une nouvelle fonctionnalité : la recherche fédérée. Elle permet de chercher simultanément dans plusieurs bases de données documentaires. La requête de l'usager est envoyée à chacune des bases sélectionnées et les résultats provenant de toutes les bases de données sont présentés dans une seule et même interface, celle du portail. En plus des bases de données bibliographiques et plein texte, on peut envoyer des requêtes à des catalogues de bibliothèques, de musées, des dépôts institutionnels ou toute autre ressource similaire.

1.1.4 Bibliothèque numérique

L'on peut définir cette dernière comme une collection organisée de documents électroniques en accès libre et généralement gratuit sur Internet, associée à une interface permettant la recherche et la consultation de ces documents. Les bibliothèques numériques sont très variables en volume et types de documents. Exemples de bibliothèques numériques : Gallica, Patrimoine numérique (catalogue des collections numérisées en France) , la Bibliothèque numérique mondiale de l'Unesco, Europeana, etc...

Que l'on parle de bibliothèque numérique ou de catalogues de bibliothèques en ligne (le Sudoc ou le Ccfr[1]), on se réfère à un portail documentaire.

1.1.5 Recherche unifiée et fédérée

On peut faire une distinction entre la recherche unifiée/intégrée et la recherche fédérée. La première interroge les collections internes d'une institution. La seconde interroge des ressources externes. Dans le cadre de ce mémoire, la distinction ne sera pas faite.

[1] Sudoc : Le catalogue du Système Universitaire de Documentation ; Ccfr : catalogue collectif de France

1.2 Avantages et inconvénients de la recherche fédérée

1.2.1 Une interface de recherche simplifiée ?

Comme le souligne Franck Cervone, dans son article *Federated searching : today, tomorrow and the future*, daté de mars 2007, « on a l'impression que la recherche fédérée est là depuis toujours, mais c'est seulement depuis trois ans qu'elle occupe le devant de la scène des bibliothèques ».

Le logiciel de recherche fédérée a été mis en place pour des raisons simples : l'internaute, lorsqu'il a de multiples bases à disposition, préfère utiliser une interface unique plutôt que d'avoir à apprendre à manipuler différentes interfaces pour chaque base. Il n'a plus à se familiariser avec les interfaces natives des différents systèmes qu'il veut consulter. « Il contrôle ainsi l'interrogation de ces systèmes par l'intermédiaire d'une seule interface » [4, Arsenault].

Des études d'usage mettent l'accent sur le besoin d'interfaces de recherche simples. « L'utilisateur ne veut pas d'interfaces de recherches qui fournissent une myriade de choix ou exigent d'eux qu'ils prennent des décisions » [5, Cervone]. La recherche fédérée permet, dans une certaine mesure, de contourner le problème du choix des ressources à interroger, pourtant étape cruciale du processus de la recherche documentaire. Les ressources sont présélectionnées par le bibliothécaire professionnel responsable des collections [4, Arsenault].

Des études démontrent que les utilisateurs ont du mal à choisir la source appropriée, ne serait-ce que pour trouver un type de document donné. Selon l'étude Tallent[2], la ségrégation des types de documents en bases distinctes constituerait l'un des problèmes les plus frustrants pour les usagers qui ne comprennent généralement pas pourquoi il est impossible de retrouver des références d'articles de périodiques dans le catalogue de la bibliothèque.

De plus, les usagers n'utiliseront pas souvent une option de recherche avancée. Ils s'attendent même à ce que le moteur de recherche arrive à comprendre ce qu'ils veulent :
« The reality is that most of these complex or advanced options are so complex or advanced that the average person cannot really work out what it is they are supposed to do, so the functionnality goes unused ».

[2] Tallent. Metasearching in Boston College Libraries:A Case Study of User Reaction. New Library World. 2004, vol. 105, n° 1196/1197, p. 71.

Selon l'étude TALLENT[3], les fonctionnalités de type « recherche rapide » qu'offre le service de recherche fédérée sont très populaires et appréciées des utilisateurs. Cette fonction de recherche rapide consiste à imiter le principe de fonctionnement des gros moteurs de recherche de type Google. La popularité de ces outils réside dans leur simplicité d'utilisation. L'usager ne se préoccupe pas du choix des sources, du type d'information ou du format de documents.

Pourquoi contraindre l'utilisateur à choisir les bases qu'il va interroger ? N'est-ce pas lui imposer de comprendre ce qu'il est en train de faire et de maîtriser un outil complexe ? C'est la meilleure manière de l'inciter à aller voir ailleurs [11, Pour ou contre…].

On peut trouver des limites à la simplification de cette interface « à la google ». On peut imaginer d'autres solutions que le simple encart pour lancer une requête sur l'interface utilisateur. Il existe d'autres logiques d'accès à l'information qui répondent aux besoins de l'usager sans complexifier les choses :

- par type de document si cela est présenté de manière simple sous forme d'onglet ou par menu déroulant (portail d'Amazone)
- par plan de classement thématique (portail de l'INRS[4])

Il s'agit d'éviter à l'utilisateur d'interroger les bases une par une, tout en lui donnant la possibilité de faire une première sélection. Le choix des sources fait partie de la recherche documentaire. Simplifier cette étape ne signifie pas la supprimer.

1.2.2 Valorisation des ressources

Un avantage de la recherche fédérée est de pouvoir trouver des sources d'informations auparavant cachées et ainsi d'accoître la visibilité des bases de données et des ressources disponibles dans une institution. L'usager interroge des bases dont il ignore l'existence. La plupart des chercheurs sont familiarisés avec un nombre limité de bases de données liées à leur champ d'expertise. Ils recherchent rarement des bases de données extérieures à leur discipline. Or le service de recherche fédérée offert par le portail documentaire leur permettra d'accéder à de nouvelles sources d'informations (bases de données, sites web etc…) fournissant ainsi de nouvelles ressources à explorer.

[3] p. 74
[4] Institut national de recherche et de sécurité

La recherche fédérée a été une manière de répondre aux attentes des utilisateurs confrontés à une augmentation massive de ressources numériques et à l'hétérogénéité de l'information structurée et non structurée, dispersée sur le web.

Les bibliothèques contrairement aux musées et autres institutions patrimoniales ont été les premières à intégrer dans leur portail cette fonctionnalité. En effet, contrairement à l'objet d'art, les imprimés sont rarement une pièce unique et les notices d'ouvrages sont plus simples que celles du patrimoine mobilier ou immobilier puisqu'il faut parfois 99 champs documentaires pour décrire un objet dans un logiciel de gestion des collections d'œuvres. Une notice d'ouvrage va rarement au-delà d'une vingtaine de champs. Quant aux archives, autre grand domaine du patrimoine, la question des droits et de confidentialité a été un frein à leur mise en ligne et leur consultation.

Face à la concurrence des portails commerciaux et grand public, et aux moteurs de recherche fédérée comme google, yahoo !, bing, exalead, les portails documentaires des institutions patrimoniales ont dû s'adapter et proposer à l'usager un accès aux contenus simplifié, intuitif et ergonomique.

L'objectif est d'accéder à des contenus hétérogènes sans avoir à multiplier les clics ou faire une sélection préalable des ressources pour éviter à l'usager d'être confronté à un trop grand choix ; d'éviter les options de recherche avancée jugées trop compliquées par les utilisateurs mais, en même temps, proposer une première sélection (par type de documents par exemple). Il s'agit de trouver un équilibre afin de rendre la recherche fédérée simple et efficace.

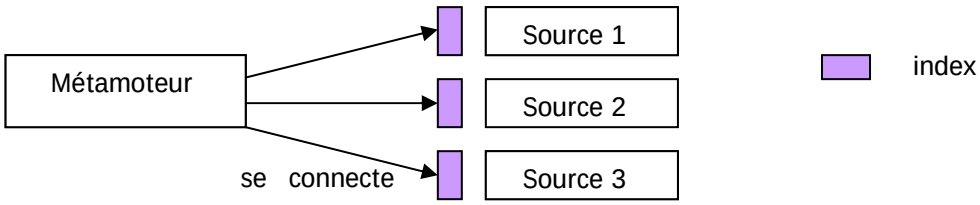
2 Outils techniques

Chaque institution culturelle a mis en œuvre ses propres systèmes informatiques adaptés aux différents types de fonds mais souvent incompatibles entre eux. Pour échanger et partager l'information entre musées, bibliothèques et archives, la mise en place de normes et standards est nécessaire.

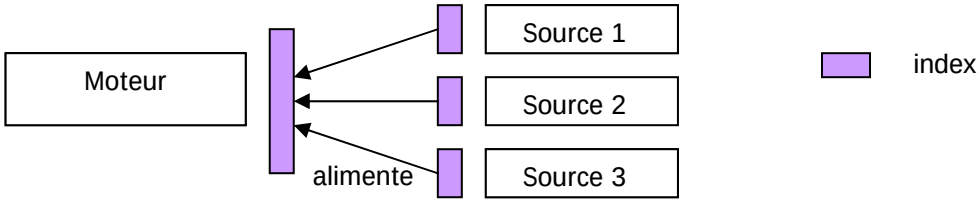
2.1 Les protocoles d'interrogation

Il existe deux grands principes et deux grands types d'outils pour la recherche fédérée : le premier fonctionne sur le principe d'extraction de données et de recherche dans un index unique via un moteur de recherche de type google ou exalead. On parle de recherche par moissonnage et de protocole asynchrone puisqu'il se fait en deux temps. Le second outil utilise des connecteurs pour traduire et envoyer simultanément la requête auprès de différentes sources via un métamoteur. On parle de recherche croisée et de protocle synchrone [7, Gibson ; 13, Foulonneau].

Recherche synchrone ou recherche croisée



Recherche asynchrone ou moissonnage



Si les sources sont structurées, la recherche fédérée peut se faire grâce aux protocoles tels que : HTTP, Z.3950, SRU/SRW, OAI-PMH. Il en existe d'autres mais nous ne les décrirons pas tous.

2.1.1 Les protocoles synchrones : exemples

Le protocole HTTP (HyperText Transfert Protocole) est utilisé dans le monde web pour accéder à des ressources multiples statiques. Pour interroger des bases de données en HTTP, il faut passer par des programmes spécifiques appelés « connecteurs » ou « wrappers », capables d'interroger des réservoirs variés. Ces programmes traduisent les questions de l'utilisateur dans un langage de requête compréhensible par le moteur de recherche de la base de données, extraient l'information pertinente, la renvoient et la traduisent « à la volée » au format compréhensible par le navigateur HTML.

Le protocole Z.3950 repose sur le principe client-serveur. Tout client normalisé Z.3950 interroge tout serveur normalisé Z.3950. Le client Z.3950 s'adresse à un ou plusieurs serveurs Z.3950, simultanément ou individuellement. Il intègre des fonctions qui permettent d'élaborer une requête structurée sur des champs spécifiques avec opérateurs booléens et des fonctions de tri de résultats. Les échanges n'utilisent pas l'URL pour transmettre les requêtes et rapatrier les données mais un langage propre, moins « visible ». Ce standard de recherches est utilisé dans le milieu des bibliothèques[5]. Il est petit à petit remplacé par le SRU/SRW plus adapté à l'univers web dans la mesure où celui-ci utilise l'adresse URL pour rapatrier l'information.

Les protocoles SRU (Search/Retrieve via URL) et SRW (Search/Retrieve Web Services) suivent le modèle fonctionnel du Z.3950 mais prennent appui sur l'infrastructure web. Maintenus par la Bibliothèque du Congrès, et destinés à faciliter la recherche sur Internet, ils se fondent sur le langage CQL (Common Query Language, standard de syntaxe de requête), qui présente l'avantage d'une (relative) lisibilité. SRU/SRW spécifient tout à la fois la syntaxe des requêtes, et celle des réponses données à ces requêtes. SRU/SRW permettent d'interroger simplement, via le protocole http du web, des bases de données jusqu'alors cachées ou plus difficilement accessibles. La version 1.1 de SRU a été publiée en février 2004, et devrait être suivie d'une version 1.2 actuellement en cours de développement. La réponse à une requête SRU/SRW est un fichier XML[6], exploitable comme tel. Le projet de La Bibliothèque Européenne TEL utilise ce protocole.

[5] La recherche fédérée de la bibliothèque Kandinsky utilise ce protocole ainsi que le SUDOC, le catalogue BN-Opale Plus de la BNF et le CCfr.

[6] Extensible Markup Language

2.1.2 Le protocole asynchrone : OAI-PMH

L'OAI-PMH est un standard de transfert de métadonnées et non de recherche à proprement parler. Cependant, il est utilisé pour la recherche fédérée[7]. L'Open Archive Initiative Protocol for Metadata Harvesting" (OAI-PMH) ou « Protocole de Collecte de Métadonnées de l'Initiative Archives Ouvertes » a été conçu à l'issue de la Convention de Santa Fé en 1999, pour implanter des bases interopérables de pré-publications scientifiques.

Ce protocole permet :

- de centraliser les métadonnées référençant diverses ressources mais laisse ces ressources à leur emplacement initial
- de mettre à jour simplement et automatiquement des métadonnées collectées et des liens, en répercutant les dernières modifications des réservoirs sources, sans copier à nouveau l'intégralité des données (la charge n'en étant que plus légère pour les serveurs) ;
- d'encourager l'utilisation d'un format de description assez générique pour les besoins transdisciplinaires, même les plus simples, sans interdire des spécifications adaptées à des besoins plus spécialisés ;
- d'intégrer, de ce fait, des ressources d'origines diverses, dans des traditions descriptives propres, sans empêcher le maintien parallèle de ces traditions pour d'autres usages.
- d'abattre des barrières du " web invisible " en rendant possible le signalement de ressources non accessibles aux moteurs de recherche

Le protocole OAI s'appuie sur quelques concepts documentaires simples [17, Nawrocki] :

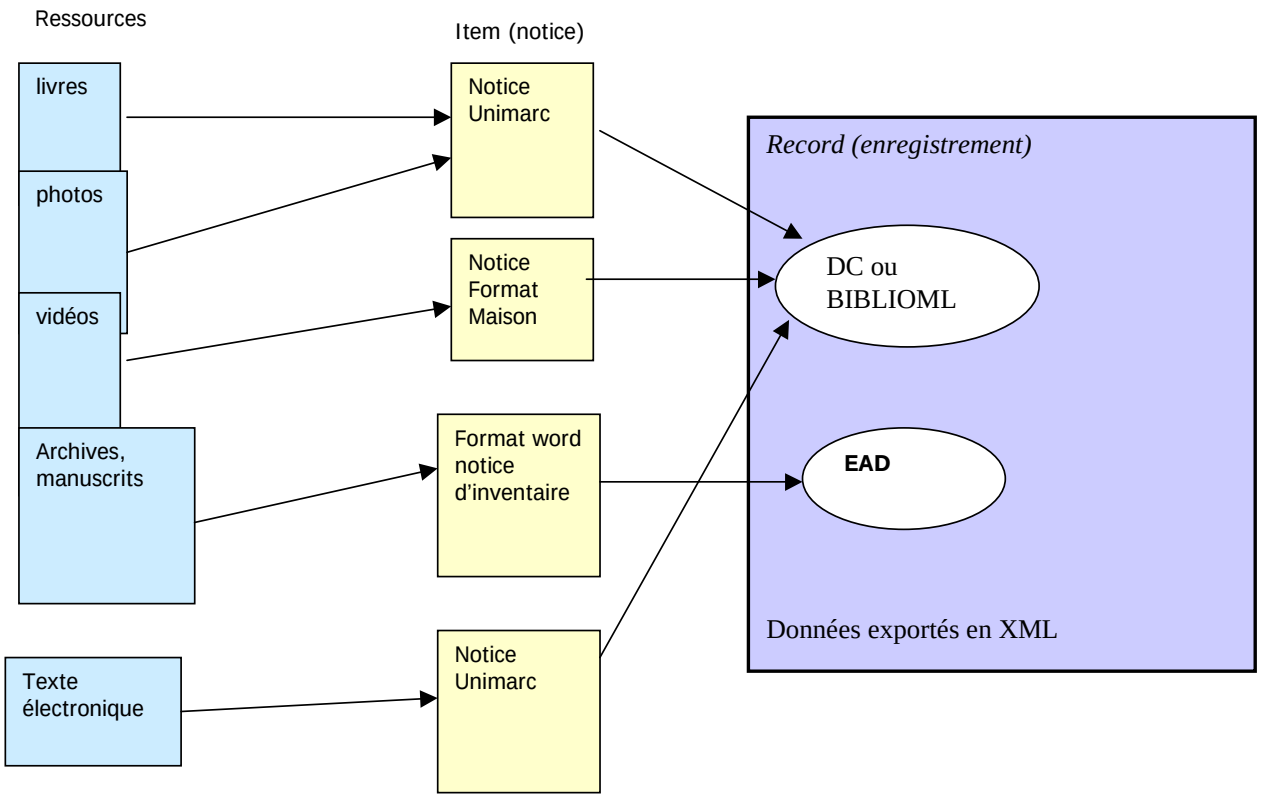
-*la ressource* : l'objet réel. Par exemple : un livre, un CD, une vidéo, une liasse de manuscrits, une image fixe. On entend évidemment par « objet réel » une image numérique ou un texte électronique. Il s'agit du document qui va être décrit ; celui pouvant être numérique.

-*l'item* : la notice informatique décrivant cet objet (exemple : une notice bibliographique au format UNIMARC)

[7] Dans son ouvrage *Le catalogue de la bibliothèque à l'heure du web 2.0 : étude des opacs de nouvelle génération*, Marc Maisonneuve ne considère pas le protocole OAI comme un protocole d'interrogation de recherche fédérée. Nous ne prendrons donc pas en compte la définition de Marc Maisonneuve dans le cadre de ce mémoire.

-l'enregistrement (record) : l'ensemble de métadonnées extraites d'un item dans un format XML, et qui fait l'objet de l'échange entre l'entrepôt et le moissonneur (exemple : la description du livre en format BiblioML ou DC[8]) ; il y a autant d'enregistrements possibles par item que de formats dans lesquels l'entrepôt est moissonnable ; une notice Unimarc peut être enregistrée en BiblioML mais aussi en Dublin Core.

-de manière optionnelle, chacun des items peut relever d'un ou de plusieurs ensembles ou lots (sets), définis par le producteur de l'entrepôt pour permettre une moisson " en bloc " de la totalité des items relatifs à un type de support ou à un thème particulier (par exemple les périodiques, l'Histoire de l'Alsace ou la division 320:sciences politiques de la classification Dewey).



Protocole OAI-PMH. Schéma inspiré de celui de François Nawrocki [17].

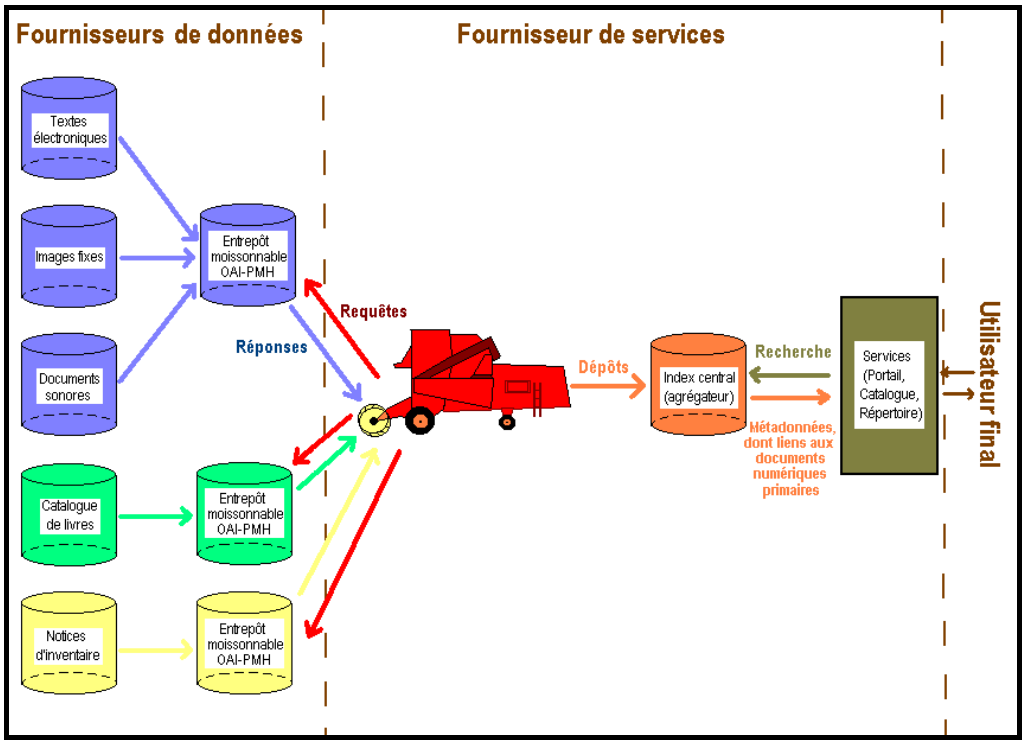
[8] Dublin Core

Dans ce contexte, le protocole définit le langage par lequel communiquent le fournisseur de données (entrepôt) et le fournisseur de services (agrégateur), qui rassemble des données collectées par un moissonneur.

Pour alimenter un agrégateur, le moissonneur visite plusieurs entrepôts, qu'il doit interroger massivement une seule fois ou en plusieurs étapes, pour extraire les enregistrements des items qui l'intéressent.

Après la collecte, le moissonneur dépose les données dans une base que le fournisseur de services rend accessible à ses clients. L'interrogation de cette base est directe et ne sollicite pas les entrepôts d'origine. En effet, l'utilisateur final interroge uniquement le réservoir de notices, constitué par moisson, du fournisseur de service, qui lui retourne en réponse la liste des notices pertinentes.

Ce notices proposent notamment un lien hypertexte vers le document primaire, seulement accessible sur le serveur du fournisseur de données : en activant ce lien, l'utilisateur arrive sur le site et dans l'environnement graphique de l'institution productrice, dont le serveur n'est finalement sollicité que pour la fourniture de ce seul document.



Architecture fonctionnelle de services OAI. François Nawrocki [17].

L'OAI-PMH et la norme Z39-50

Au contraire de portails fédérateurs de ressources ou de catalogues collectifs reposant sur la norme d'interrogation Z39-50, une base de donnée constituée par moissons OAI permet au fournisseur de service de rendre accessibles des données descriptives de documents sans faire peser une lourde charge sur le serveur de chaque fournisseur de données ; de plus, le temps de réponse au client final dépend du seul serveur du fournisseur de service (et non du serveur le moins performant de l'ensemble des fournisseurs de données, comme c'est le cas dans une architecture Z39-50).
En revanche, les données exploitées par le fournisseur de service sont le reflet d'un état figé des données collectées, en date du dernier passage du moissonneur, alors qu'une interrogation simultanée de plusieurs bases par transfert Z39-50 permet d'afficher les données en temps réel. Par conséquent, l'OAI-PMH n'est pas toujours la solution organisationnelle et technique la plus pertinente pour des entrepôts dont une part importante du contenu est soumis à des modifications très fréquentes, par exemple quotidiennes (actualités, œuvres vivantes…).

2.2 Echange de données de systèmes hétérogènes

L'échange de données consiste à faire communiquer de l'information entre différents systèmes plus ou moins hétérogènes.

La notion de métadonnée est utilisée dans le contexte des systèmes d'information moderne remplaçant le terme de catalogage. La difficulté consiste à utiliser un système de vocabulaire contrôlé global car même si des éléments communs de métadonnées sont utilisés, le contenu de ces éléments n'est pas forcément compatible [18, YOUSEFI].

2.2.1 Les métadonnées

Les métadonnées peuvent être définies comme un ensemble structuré de données créées pour fournir des informations sur les ressources électroniques. Les métadonnées peuvent être [38, Raïs]:
-Descriptives : description et identification des ressources : titre, source, date, volume
-Administratives : gestion et conservation des documents : droit d'utilisation, droit d'auteur, cycle de vie, contrôle de qualité

-Structurelles : pour la navigation et la présentation ; elles permettent d'établir des liens entre les documents, partie constituante du web sémantique : titre de page, table des matières, chapitres, parties, index, relations entre les composants de la ressource.

Les métadonnées peuvent être décrites actuellement selon plusieurs standards : RDF[9], TEI[10], syntaxe Meta HTML, Dublin Core, EAD[11] etc…

Si la notion de métadonnées s'apparente au travail classique de traitement documentaire réalisé depuis longtemps par les bibliothécaires et documentalistes, on l'utilise généralement pour les ressources du web. Elles ne sont pas restreintes à la description catalographique ; il existe des catégories supplémentaires de métadonnées pour gérer les ressources électroniques, les liens avec d'autres documents, la navigation, le contrôle des accès, la gestion du cycle de vie etc… Elles peuvent concerner :
-Un ensemble de ressources : un site web
-Une ressource individuelle : une page d'un site web
-Une partie d'une ressource : une photo sur une page web

L'intérêt des métadonnées se situe au niveau de la description des ressources : description du contenu avec la possibilité de multiplier les synonymies pour faciliter la recherche et d'exploiter les collections numériques (description des relations entre les fichiers).

C'est aussi la possibilité pour le moteur de recherche d'extraire automatiquement l'information structurée sur le document, repérer des documents non textuels qui seraient invisibles à retrouver sans les métadonnées (images, documents audiovisuels…). Les métadonnées peuvent être gérées automatiquement par le système ou de manière manuelle. Elles peuvent être constituées de contenu structuré (Auteur, date, mot-clés) ou non (Titre, Description…).

Elles peuvent être externes, c'est-à-dire contenues dans une notice séparée du document : notice dans un catalogue de bibliothèque avec référence à la ressource, notice stockée dans une base de données spécifique avec un lien pointant vers la ressource.

Ou internes : elles sont intégrées dans la notice elle-même

[9] Resource Description Framework
[10] Text Encoding Initiative
[11] Encoded Archival Description, en français *Description archivistique encodée*

Englobantes : la ressource elle-même est balisée et les métadonnées englobent le contenu du document (ex : EAD, TEI).
Encapsulées : les métadonnées sont incluses dans le document (ex : balises META dans un document HTML).

Elles peuvent être généralistes :
Dublin Core : description des ressources électroniques
MARC : description des ouvrages avec différents standards, MARC21, UNIMARC
ONIX : passerelle entre éditeurs et bibliothèques

Les métadonnées peuvent être spécialisées. Exemples de métadonnées par type de ressources :
Archives : EAD, EAC[12]
Manuscrits : MASTER[13]
Texte : TEI
Thèses : TEF[14]
Images : IPTC[15], XMP[16]
Ressources audiovisuelles : MPEG-4, MPEG-7[17]
Ressources muséographiques : CIMI XML Schema for spectrum[18]
Exemples de métadonnées par domaines :
-pédagogie : LOM[19], pour la description des ressources liées à l'éducation
-sciences sociales : DDI[20]

Elles peuvent être informatiques :
Les balises <Title> et <meta> des pages HTML
Les champs IPTC des images JPEG/TIFF
Les champs EXIF[21] des images JPEG
Les propriétés des documents MS Office (excel, word…)

---

[12] EAC : Encoded Archival Context
[13] Manuscript Access through
[14] Thèses électroniques françaises
[15] International Press Telecommunications Council. Structure de fichier et ensemble de métadonnées créés afin d'accélérer les échanges internationaux de nouvelles parmi les journaux et les agences de nouvelles.
[16] eXtensible Metadata Platform
[17] Multimedia Content Description Interface.
[18] Computer Interchange of Museum Information. Standard procedures for Collections Recording Usec in Museums.
[19] Learning Object Metadata
[20] Data Documentation Initiative

Pour que les métadonnées puissent remplir leur objectif, elles doivent être interopérables. Pour cela, il est nécessaire d'utiliser des métalangages, établir des équivalences entre syntaxes et terminologies, utiliser des protocoles d'interrogation et d'échanges.

Le métalangage XML permet d'aboutir à une seule représentation de l'information, de faire communiquer entre elles des bases hétérogènes et de les interroger simultanément.

2.2.2 XML, un format d'échanges

La norme XML est un outil permettant de définir un métalangage, c'est-à-dire de créer des documents structurés en définissant le vocabulaire et la syntaxe de ces données. XML utilise des balises sémantiques qui permettent de donner un sens au contenu qu'elles renferment. Les balises du HTML sont des balises de présentation et non de contenu. Elles structurent une page web à l'aide de codes. Le navigateur lit et interprète : telle valeur doit s'afficher de telle manière. HTML ne distingue pas le fond de la forme.

Un fichier XML doit être accompagné de la déclaration de sa structure appelée DTD[22]. Le parseur, logiciel qui lit les données XML, vérifie si celles-ci sont bien conformes aux règles définies dans la DTD. Si un attribut Y est défini comme obligatoire dans la DTD mais ne se trouve pas dans le fichier XML, celui-ci n'est pas valide. La DTD n'étant pas écrit en langage XML, elle évolue vers le schéma, document écrit en XML. Ce qui permet d'avoir un seul langage.

Le principe du XML est de pouvoir écrire notre document comme on en a envie. L'intitulé des balises, leur imbrication, leur caractère obligatoire ou facultatif, leur ordre de succession, si elles ont plusieurs valeurs… Toutes ces règles sont déclarées dans la DTD ou le schéma.

Les feuilles de style CSS[23] ou XSLT[24] affichent les données selon les modèles voulus.

L'échange de données consiste à faire communiquer de l'information entre différents systèmes plus ou moins hétérogènes. Par exemple, deux bases de données gérées par deux outils différents. La norme XML est utilisée comme le principal format d'échange de données entre systèmes hétérogènes.

---

[21] Exchangeable Image File Format
[22] *Définition de Type de Document*
[23] *Cascading Style Sheets* : feuilles de style en cascade
[24] *eXtensible Stylesheet Language Transformations*

Le problème de l'intégration de sources hétérogènes de données en une interface commune ne consiste pas uniquement à résoudre le problème du protocole d'échanges et d'homogénéisation des formats documentaires, il consiste aussi à exprimer les relations entre les données des sources intégrées. La recherche fédérée pose le problème du « bon mot-clé » quand il n'existe pas de thésaurus transversal permettant l'interrogation de l'ensemble des catalogues. L'idéal d'un système d'information est de pouvoir retrouver, si j'interroge avec le mot-clé « rituel du mariage », tous les documents sur ce sujet. Or, ces documents ne sont pas nécessairement indexés avec ce descripteur « rituel de mariage » puisqu'ils appartiennent à des bases de données différentes utilisant leur propre langage contrôlé. La recherche d'information ne sera alors pas performante.

L'interopérabilité est la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre[25]. En d'autres termes, l'interopérabilité est la capacité à dialoguer, à interagir, et à échanger des données (avec le minimum de perte d'information et de fonctionnalités) de deux systèmes disposant de différentes caractéristiques en terme de matériels, logiciels, structures de données et interfaces.

Donner accès à plusieurs sources de données en une interface unique suppose l'interopérabilité des bases de données et des outils d'indexation communs. Mais quand il n'existe pas de langage d'indexation commun, qu'en est-il l'interopérabilité sémantique[26] ? L'interopérabilité sémantique est une réponse à l'hétérogénéité sémantique des informations traitées par les diverses applications. Elle implique que les différents utilisateurs partagent des vues cohérentes sur les systèmes de concepts propres aux diverses applications.

---

[25] Définition du Groupe de travail de l'AFUL (Association Francophone des Utilisateurs de Logiciels Libres).

[26] On distingue plusieurs types d'interopérabilité dont l'intéropérabilté technique, sémantique et syntaxique

- Interopérabilité technique

L'interopérabilité technique permet à des systèmes de communiquer grâce à des protocoles et langages similaires ou pour lesquels il existe une procédure d'équivalence.

- Interopérabilité sémantique et syntaxique

L'interopérabilité sémantique est possible lorsque les métadonnées sont similaires ou comprennent des liens d'équivalences car elles représentent les mêmes concepts.
L'interopérabilité syntaxique suppose que les métadonnées ont une syntaxe similaire ou qu'une procédure d'équivalence existe. Par exemple, une date peut être encodée de la manière suivante : « 28-01-75 » ou « 28 janvier 1975 » ou encore « 1975-Jan-28 ».

Les institutions culturelles proposent au grand public de consulter ses collections en ligne. A titre d'exemple, Europeana, la bibliothèque numérique européenne, donne à l'internaute la possibilité en une seule requête d'effectuer des recherches et de naviguer dans les collections numérisées des bibliothèques, des archives et des musées européens.

Chercheurs, spécialistes et grand public ont ainsi accès aux différents catalogues du Musée du Quai Branly, de la Cité de l'Histoire de l'immigration, de la Bibliothèque Kandinsky….etc. Ils peuvent interroger, par le critère de recherche « mot-clé », l'ensemble des ressources hétérogènes : archives, objets, image fixe et animée, son, imprimés. Ainsi, la recherche fédérée permet d'explorer des sujets sans avoir à rechercher et à consulter chaque source séparément. Comment ces institutions patrimoniales ont-elles mis en place leur portail de recherche fédérée ? Quel protocole, quelles métadonnées, quels formats documentaires, quels champs de recherche ont été choisis ? Ont-elles réussi à établir des équivalences de vocabulaire, les divers langages documentaires ont-ils été harmonisés et si c'est le cas, comment ?

Il s'agit de comprendre, à partir de plusieurs exemples (Europeana, le Musée du Quai Branly, le portail Collections…), comment l'indexation a été construite pour un accès transversal grand public.

Deuxième partie
Portails documentaires : quelles solutions d'harmonisation?

1 Méthode d'enquête

1.1 Les critères de sélection des interfaces d'interrogation étudiés

Les portails étudiés et présentés sont au nombre de cinq : **Europeana[27], La Cité de l'histoire de l'immigration[28]**, **la Bibliothèque Kandinsky[29], Collections[30] du Ministère de la Culture et le Musée du quai Branly[31].**

Ces portails documentaires appartiennent tous à des institutions culturelles. Ils donnent accès à des ressources hétérogènes, c'est-à-dire à différents types de support : imprimés, iconographie (photos, estampes, cartes, affiche…), vidéos, sons, manuscrits, objets de collections muséales. Sites web et bases de données externes peuvent aussi être des sources interrogeables.

Les portails des musées d'ethnographie et d'anthropologie tels que le Musée du quai Branly ou la Cité de l'histoire de l'immigration (Cnhi) ont été sélectionnés car ils se rapprochent le plus du futur portail documentaire du MuCEM, point de départ de cette enquête (cf. partie III). De plus, ces portails possèdent une base « objets » et une base « archives », deux supports dont le vocabulaire d'indexation et les champs documentaires restent problématiques à harmoniser.

Le portail *Collections* a été choisi car il présente, d'une part, une recherche multi-sources exhaustive : archives, objets, imprimés et audiovisuel. D'autre part, il ne propose pas de critère de recherche spécifique « mot-clé ». La recherche se fait en langage naturel, à l'aide d'une interface simplifiée : un unique encart à partir duquel l'internaute lance sa requête.
Le portail de la Bibliothèque Kandinsky a été retenu pour sa fonctionnalité de recherche fédérée sur des bases externes. Il propose un accès à ces bases par critère de recherche « sujet ».

---

[27] http://www.europeana.eu
[28] http://www.histoire-immigration.fr/index.php?lg=fr&nav=997&flash=0
[29] http://bibliothequekandinsky.centrepompidou.fr/clientBookline/toolkit/P_Requests/formulaire.asp?INSTANCE=INCIPIO&GRILLE=MULTICRITERE_0
[30] http://www.culture.fr/fr/sections/themes/collections?typeSearch=collection&SearchableText=&SearchWhere=

Le fait que l'interface d'interrogation possède un champ de recherche « sujet/mot-clé » a été un critère de sélection, sauf pour le portail *Collections*, cas particulier à étudier.

1.2 Les personnes-ressources

Il a été convenu de rencontrer et d'interroger les personnes responsables des portails documentaires précédemment cités pour les interroger sur les méthodes retenues tant du point de vue informatique que documentaire : quels outils et quels normes ont été mis en place pour faire fonctionner la recherche fédérée ?

Au départ, la question de l'affichage des résultats a été abordée de façon sommaire puisqu'il s'agissait de savoir comment les différentes institutions étudiées avaient rendu accessible leurs catalogues par critère de recherche « sujet ». Au cours des entretiens et évaluations des portails documentaires, cet aspect s'est avéré important.

La question d'un moteur de recherche sémantique a rapidement émergé ouvrant le champ aux questions du traitement automatique de la langue (TAL), du web sémantique et à la création d'ontologie. C'est pourquoi ces aspects sont abordés dans la troisième partie.

[31] http://www.quaibranly.fr/cc/pod/recherche.aspx?b=5&t=1

2 Présentation des portails étudiés

2.1 Portails ayant harmonisé leurs mots-clés

2.1.1 Cité nationale de l'histoire de l'immigration (Cnhi)

La médiathèque Abdelmalek Sayad est ouverte depuis avril 2009. Son portail documentaire donne accès au catalogue de la médiathèque qui propose la consultation d'imprimés, de documentation sonore (cassette) et audiovisuel (DVD, cassette), de fonds iconographiques (photo, affiche). Pour le moment, il existe 25.000 notices bibliographiques.

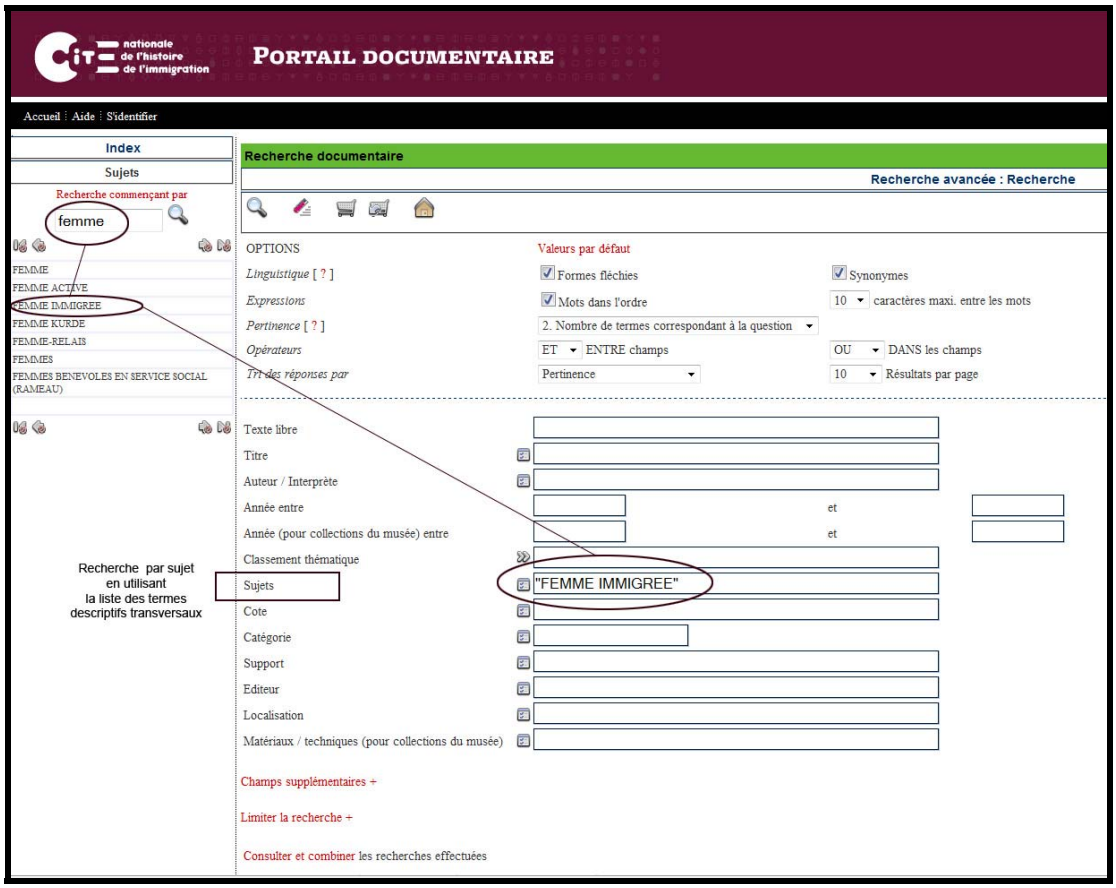
80 personnes travaillent à la Cité dont 10 à la médiathèque (7 documentalistes et 3 magasiniers). Le fonds existait déjà et venait du centre de ressource de l'ADRI[32].

La plate-forme utilisée est CADIC-intégrale. Il n'y a pas de recherche fédérée mais une seule base avec plusieurs masques de saisie selon le type de support : objets, archives, imprimés, audiovisuel, images fixes….

Le format documentaire est en fonction du type de support à cataloguer. Des bases de données Filemaker sont aussi utilisées. Elles sont importées dans Cadic (20,000 notices), grâce à un système de passerelle Xml.

La recherche est simple (full-text), avancée (titre, auteur, n° inventaire…etc et par **sujet**) ou thématique.

[32] Agence pour le développement des relations interculturelles. Elle gère un centre de ressources documentaires à vocation nationale sur l'intégration des communautés immigrées en France et la politique de la ville.



Interface de recherche du portail de la Cnhi. Recherche par mot-clé « femme ».

⇒ Il existe un thesaurus transversal. Le langage d'indexation est commun aux différents types de support. Il a été construit à partir d'un plan de classement, du thesaurus de l'ADRI évolué, de la classification Dewey et de listes diverses. L'harmonisation des mots-clés a été faite ainsi.

Un projet de recherche fédérée est en cours comprenant la base locale et des bases externes partenaires. Les bases objet et archives n'ont pas encore été intégrées à CADIC-intégrale. Cela devrait être fait prochainement mais il ne s'agit pas de recherche fédérée comme cela a été expliqué.

2.1.2 Musée du quai Branly

Le portail documentaire met en ligne quatre catalogues :

-catalogue des objets

-catalogue de l'iconothèque (photo, affiches, cartes postales, gravure, dessins)

-catalogue de la documentation muséale et des archives (dossier documentaire, inventaire, fonds d'archives privés ou publics)

-catalogue de la médiathèque : imprimés (ouvrages, périodiques, usuels, catalogues de vente, catalogues d'expositions, thèses, cartes géographiques), documents audiovisuels et documents sonores.

Toutes ces références sont issues du musée de l'Homme, de l'ancien Musée national des arts d'Afrique et d'Océanie, auxquels s'ajoutent les nouvelles acquisitions du musée du quai Branly, les dons de collections de spécialistes (fonds Condominas, Girard, Kerchache, Nesterenko) et les nouvelles acquisitions.

La recherche fédérée se fait uniquement sur les catalogues des objets, de la documentation muséale et de l'iconothèque. Le catalogue de la médiathèque [imprimés, documents sonores et audiovisuels] est interrogeable à part. Le protocole OAI-PMH n'est pas utilisé. L'interrogation des différentes bases se fait par export des contenus et correspondance des champs.

Le musée du quai Branly utilise le logiciel TMS[33] pour la base de donnée Objets (créée en 2001) et la base de données iconographique (créé en 2005 : photos et arts graphiques) [50, Guillot]. Pour les archives, il s'agit d'une base php-Mysql. La médiathèque est sur une plate-forme Flora de la société Ever. Le moteur de recherches est *Collection connexion*. Il n'y a pas de recherche simultanée sur l'ensemble de ces bases : objet, iconothèque, documentation muséale, imprimés, audiovisuel.

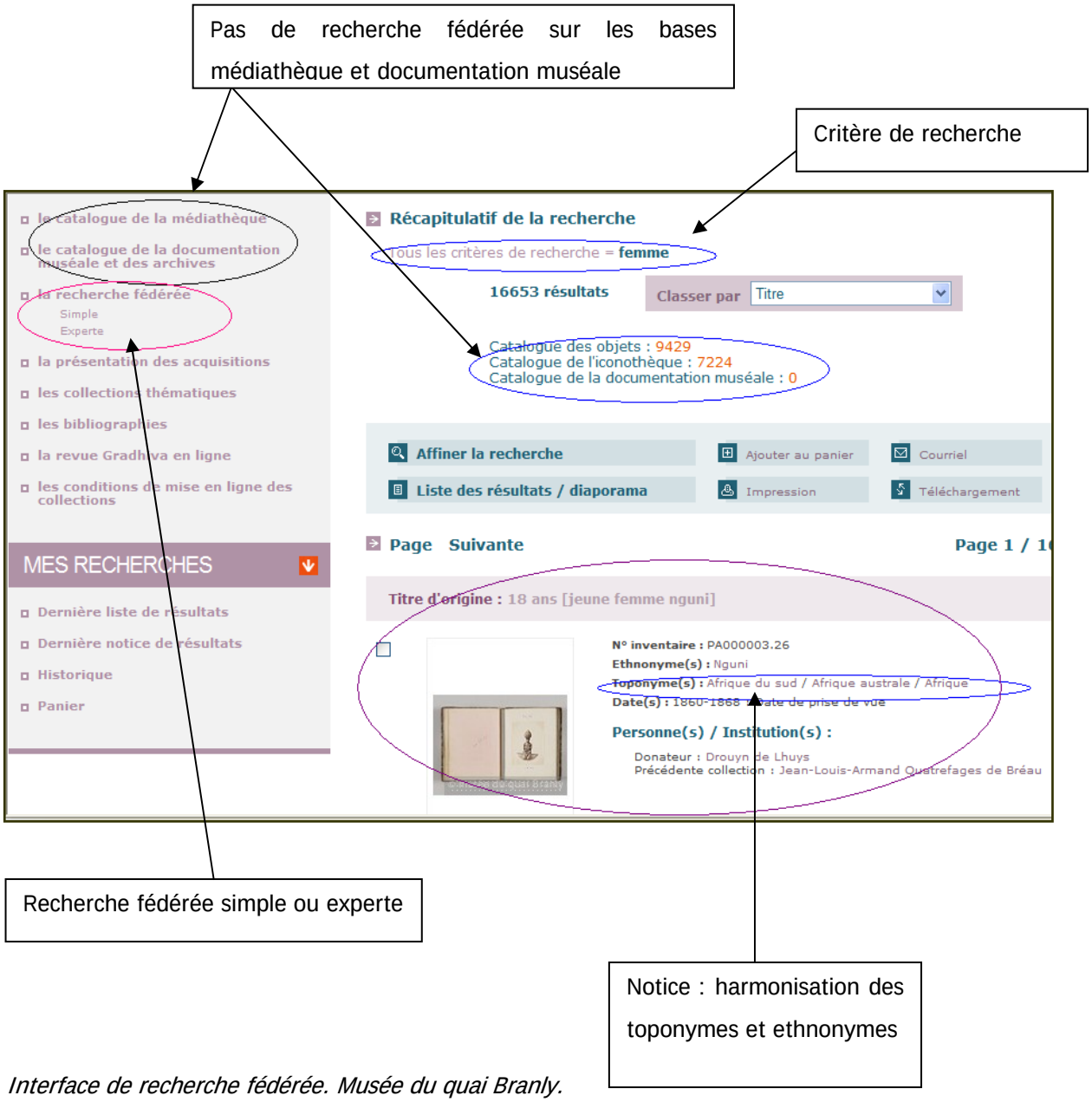
⇒ Concernant la médiathèque, l'harmonisation des mots-clés a été faite avec le langage contrôlé Rameau[34]. L'indexation transversale concerne les références de la médiathèque uniquement (imprimés, documents sonores et audiovisuels).

Concernant la recherche fédérée des catalogues objets/iconothèque/documentation muséale et archives, elle peut être simple et experte. Les critères de recherche sont : Titre, N° inventaire, Toponyme, Ethnonyme, Institution, Date, Collection.
⇒ Il n'y a pas de requête par *sujet* pour la recherche fédérée. Le « sujet » est donc demandé par le critère « mot du titre ». Mais une harmonisation des « toponymes » et « ethnonyme » a été faite. Un thesaurus transversal a été réalisé pour les critères d'interrogation *lieu* et *ethnie*, en «croisant» le vocabulaire Rameau et des mots libres déjà existants.

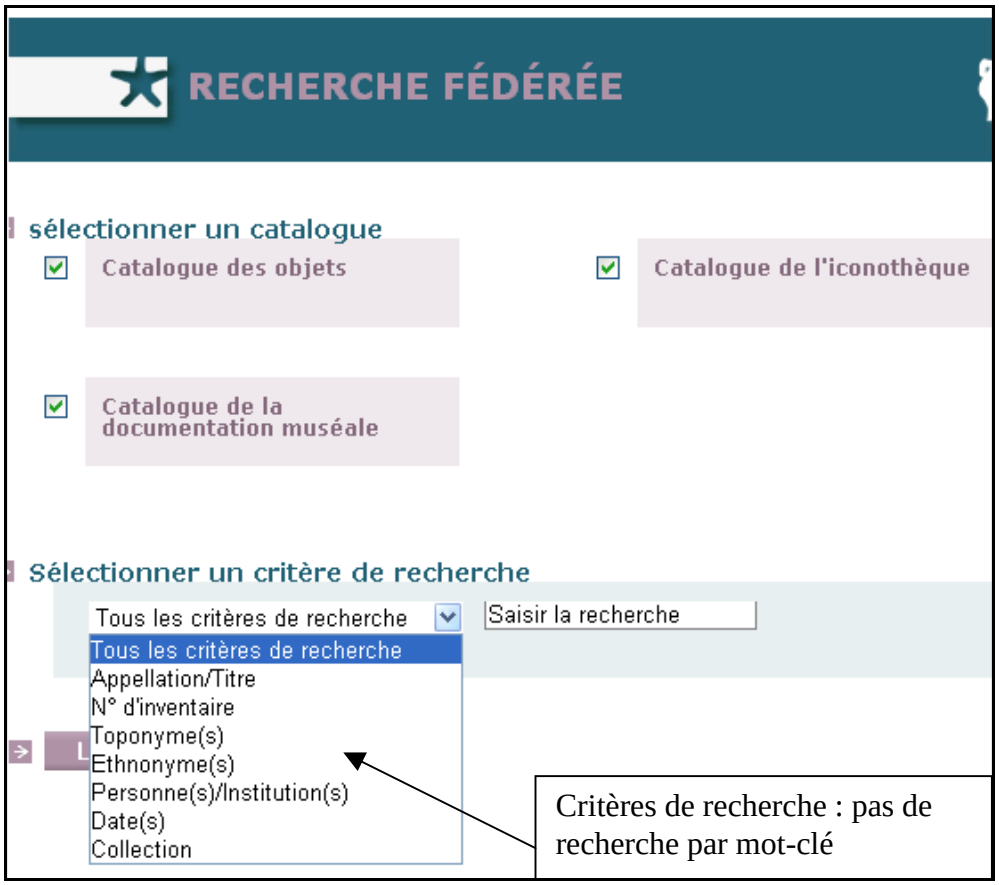
---

[33] The Museum System
[34] La BPI (Bibliothèque publique d'information) a utilisé le même outil d'indexation commun pour ses collections (imprimés, site web, documents audiovisuels et sonores). Cf. annexe.

A noter : actuellement, la recherche fédérée ne se fait que sur deux catalogues : objets et iconothèque.



Interface de recherche fédérée. Musée du quai Branly.



Musée du Quai Branly. Critères de recherche.

2.2 Portails sans harmonisation des mots-clés

2.2.1 La Bibliothèque Kandinsky

La Bibliothèque Kandinsky est un centre de documentation et de recherche du Centre Georges Pompidou. À l'origine, ce service était essentiellement réservé aux conservateurs du musée. Aujourd'hui, la bibliothèque donne désormais accès à ses collections. Chercheurs et étudiants peuvent les consulter.

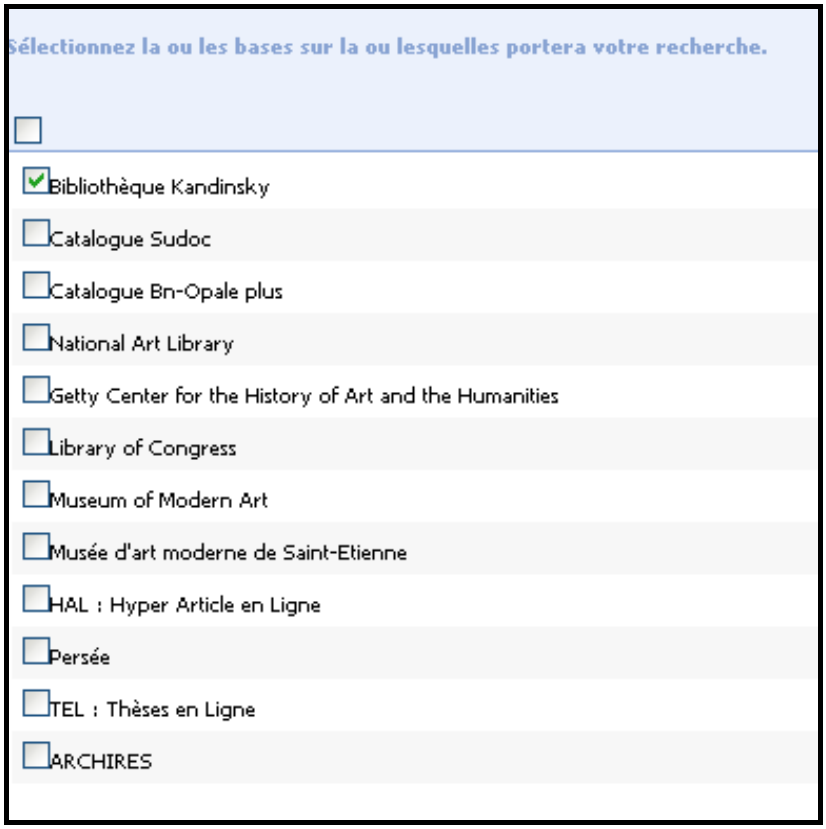
Les collections de cette bibliothèque sont destinées à la recherche et à l'exposition. Elles sont consacrées aux artistes et aux oeuvres d'art plastique, de design, d'architecture, à la photographie, au cinéma, à la vidéo et aux nouveaux medias des 20ème et 21ème siècles.

OREX est la plate-forme technique qui gère la collection des œuvres, le portail de la bibliothèque Kandinsky et les autres bases. Créé en 1999-2000, ce système d'information est en cours de changement. Une nouvelle base de données des collections doit être mise en place avec le logiciel *Videomuseum.*

Un projet de recherche fédérée est en cours avec l'idée d'un accès unique vers ces bases de données : la bibliothèque Kandinsky, des bases externes (Catalogue Sudoc, Museum of Modern Art, Getty Center …etc) et les bases internes (base collections, base manifestations et deux bases d'archives), avec de l'OAI-PMH en Xml Dublin Core et EAD.

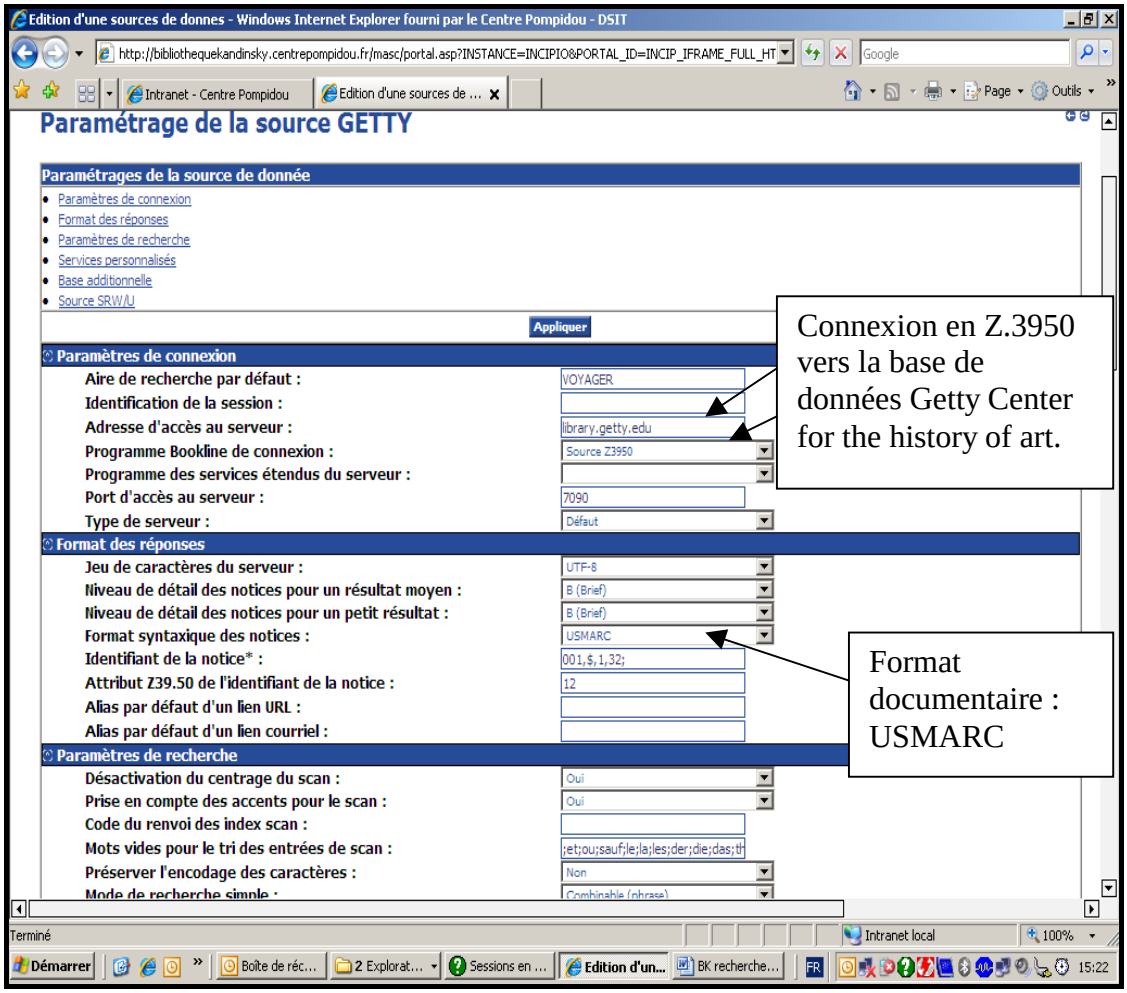
Actuellement, le portail de la bibliothèque Kandinsky permet, au-delà de l'interrogation de son catalogue interne, une interrogation simultanée des catalogues de plusieurs bibliothèques spécialisées et de ressources extérieures.

Pour cela, il faut aller sur le portail, cliquer sur le lien *Recherche avancée*, puis sélectionner les sources que l'on désire interroger : Bibliothèque du Congrès, HAL, Persée, Sudoc, Bn-opale, Getty Center for the history of art, Museum of modern Art, Musée d'art moderne de Saint-Etienne, Archire.

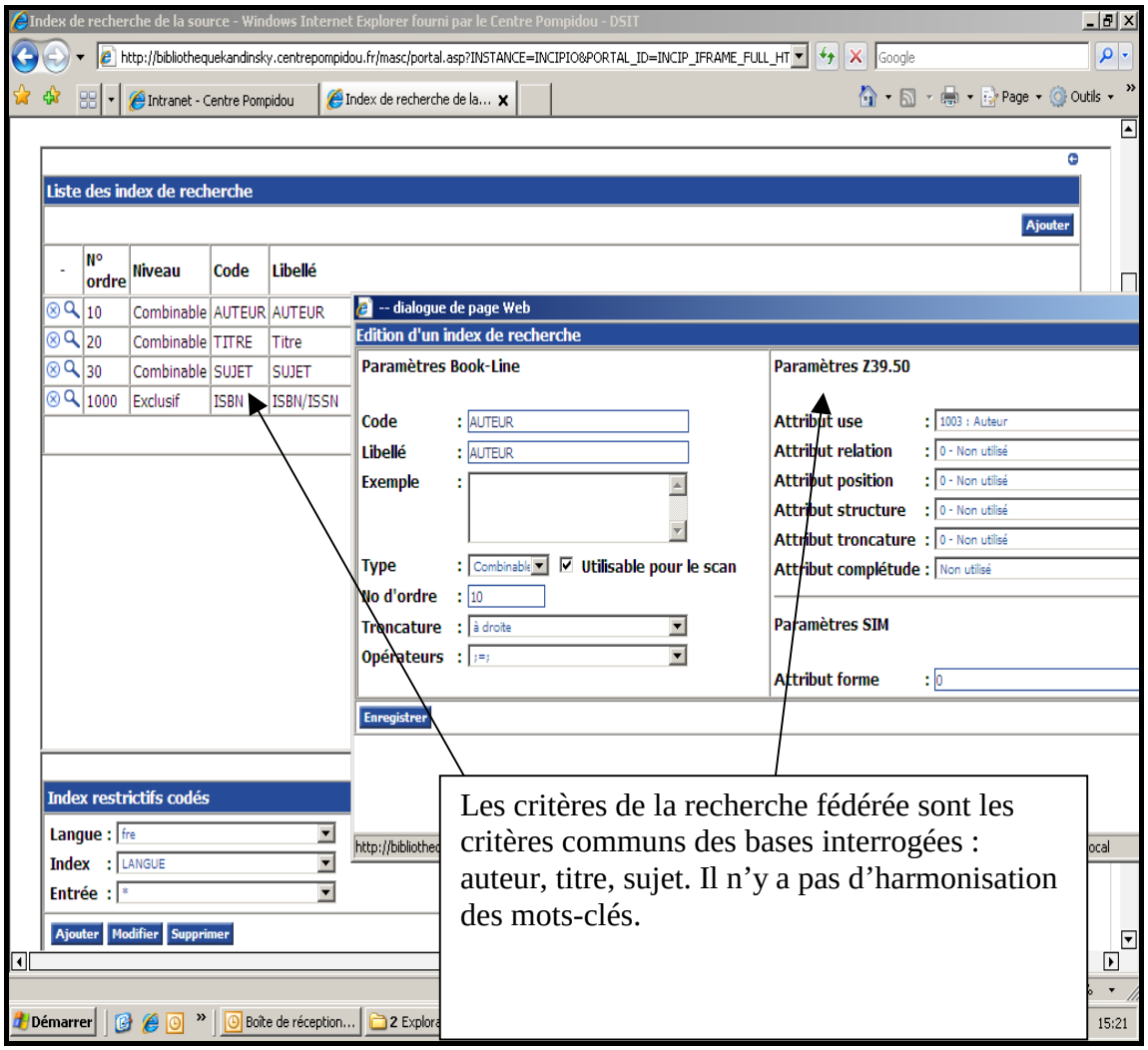


Bibliothèque Kandinsky. Sélection des bases sur lesquelles va porter la recherche fédérée.

Les protocoles d'interrogation utilisés sont : Z.3950 et OAI-PMH (Sim)



Recherche fédérée de la Bibliothèque Kandinsky. Protocle d'interrogation pour accéder à la base Getty Center for the history of art. Back-office. Exemple de protocole Z3950.



Bibliothèque Kandinsky. Critères de recherche d'interrogation des bases extérieures

⇒La recherche fédérée de la bibliothèque Kandinsky utilise les protocoles d'interrogation OAI-PMH et Z.3950. Il y a une interrogation des bases externes par critère de recherche «sujet » mais pas d'harmonisation de mots-clés.

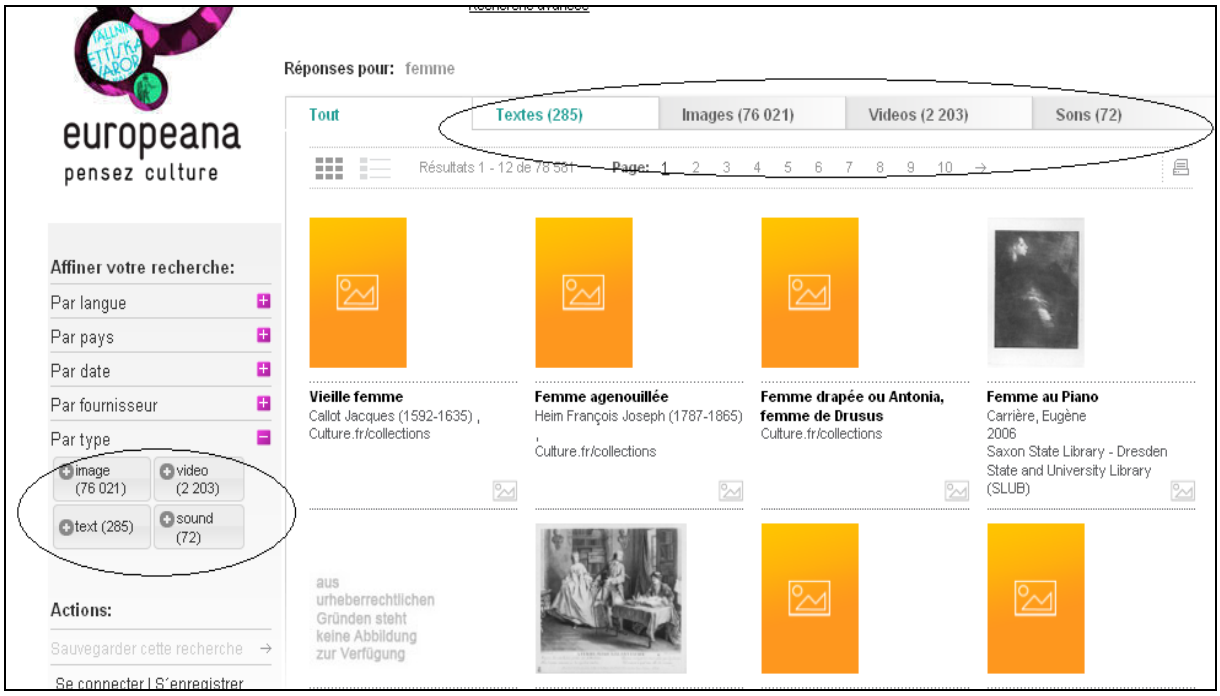
2.2.2 Europeana

Europeana est une bibliothèque numérique qui regroupe plus de quatre millions de documents libres de droit issus des collections d'une centaine d'institutions culturelles (archives, musées et surtout bibliothèques publiques) des 27 pays membres de l'Union Européenne (4 000 documents numérisés par la Bibliothèque nationale Széchényi de Hongrie, 1 000 documents numérisés par la Bibliothèque nationale du Portugal...etc). Actuellement, la France est le principal contributeur avec les documents de la Bibliothèque nationale (BNF) et l'Institut national de l'audiovisuel (INA) [3, Culture et recherche].

Les types de documents en ligne sont des images, du texte, du son et des vidéos. La recherche est simple ou avancée.

La recherche par mots est proposée tout d'abord comme une recherche simple dans un seul champ de recherche. Il est possible de choisir des fonctions de recherche avancée de type champ de données structurées (auteur, titre, date, sujet). Les données qui sont traitées pour ces fonctions de recherche sont les métadonnées descriptives, le texte des tables des matières et le texte intégral des documents textuels qui ont été numérisés en mode texte. Ces trois catégories de contenus indexés ont un poids différent pour le classement des résultats de recherche. La notice pèse plus lourd que la table des matières qui elle-même pèse plus lourd que le texte intégral [1, Lupovici].

A l'affichage, les résultats sont affichés par type de documents avec possibilité d'affiner la recherche par langue, date, source ou type de document.



Europeana. Affichage des résultats par type de support. Critère de la requête : « femme ».

Le protocole d'interrogation utilisé est l'OAI-PMH quand le partenaire institutionnel dispose d'un entrepôt OAI PMH, mais aussi FTP, des DVD. Open Search a aussi été utilisé pour récupérer les métadonnées du National Maritime Museum en Grande Bretagne. La norme utilisée pour les métadonnées est le Dublin Core. Le mode de recherche découverte simple et avancée s'appuie sur les fonctionnalités de base du free ware Lucene.

⇒ Il n'y a aucune harmonisation des mots-clés et, pour l'instant, la recherche par sujet est en mots libres, sans aide par des thesaurus, sans outil multilingue particulier bien qu'Europeana rassemble des notices des 27 pays de l'Union européenne.

2.3 Portails de recherche sémantique

2.3.1 Collections

2.3.1.1 Présentation

Le portail « Collections » donne accès à trois millions de références sur le patrimoine culturel français et parfois étranger : œuvres de musées, documents patrimoniaux des bibliothèques, fonds d'archives, patrimoine monumental et mobilier, sites archéologiques …etc. Deux millions de ces références sont illustrées par des documents numérisés [3, Culture et recherche].

Les bases de données du Ministère de la Culture ont été créées dans les années 70 avec la plate-forme Mistral. En 1998, elles sont mises en ligne et accessibles au grand public et non plus qu'aux chercheurs. En 2008, elles sont consultables simultanément via un guichet unique grâce au portail Collections. Les bases de données restent intactes. L'interrogation multi-bases enrichit la recherche [6, Collin].

Le portail interroge simultanément en langue naturelle plus de trente sources documentaires différentes (bases de données, sites internet statiques et dynamiques, publications électroniques) produites par les services du ministère de la culture et ses établissements (musée du Louvre ou du Quai Branly par exemple). Il intègre aussi les données produites par les collectivités territoriales partenaires (archives départementales, bibliothèques municipales, services régionaux de l'inventaire, musées…). La recherche est donc multi-supports : objets mobiliers et immobiliers, iconothèque (photos, estampes), manuscrits, sites, publications électroniques, imprimés. On accède aux bases de données suivantes : ARCHIM (ARCHives nationales IMages de documents), Centre historique des Archives nationales (CHAN), Centre pompidou, Agence photographique de la Réunion des musées nationaux (RMN), Mérimée, Mémoire, MediatheK, BN Opale plus, Narcisse, CCFr (BNF, SUDOC, Patrimoine), bases architecture et patrimoine, Joconde, PhoCEM (Musée des civilisations de l'Europe et de la Méditerranée), portail documentaire de la Cité de l'architecture et du patrimoine…etc

2.3.1.2 Aspects techniques du moissonnage et interopérabilité des données

Trois techniques sont utilisées afin d'attaquer les bases de données par l'intermédiaire d'un entrepôt constitué d'un index XML provenant :

-des données exportées des bases fonctionnant sous Mistral ou des sites au format HTML, le tout stocké au DSI (Département des systèmes d'information du ministère de la Culture)

-de documents en XML « natif » via la plate-forme SDX de certaines bases hébergées, elles aussi au SDI.

-de données provenant de sites externes, converties par les établissements publics concernés et récupérées par transferts FTP dans l'entrepôt [6, Culture et recherche].

Trois phases se distinguent :

- la récupération des données, transformées le cas échéant en XML et stockées dans un entrepôt. Un moissonnage sans stockage de données est possible pour certaines bases compatibles avec des connecteurs OAI (la base du Centre d'archives d'outremer par exemple)
- l'indexation des contenus base par base. Les gestionnaires des bases ont défini certains champs de leurs applications, contrôlés ou libres, comme interrogeables par le moteur *Intuition* de la société Sinequa et générateur des métadonnées *Qui, Quand, Où*.
- la mise à jour des données, récupérées par transfert ou par connecteur OAI, selon une périodicité définie selon chaque base (3, Culture et recherche).

2.3.1.3 Un moteur de recherche sémantique

Le moteur de recherche sémantique *Intuition* (Sinequa) produit des index en format XML. Une passerelle permet le moissonnage des ressources via le protocole OAI-PMH. La recherche peut être simple ou avancée. Des fonctionnalités multilingues autorisent des interrogations en anglais et en espagnol. Les résultats s'affichent en liste, en mosaïque d'images ou par catégories de documents. L'utilisateur est renvoyé directement au document source dans la base d'origine ou à des catalogues comme celui de patrimoine numérique.

La recherche se fait en plein texte, sur toutes les bases et sur tous les types de documents afin de limiter le « silence ». Le moteur de recherche est un moteur linguistique. Il gère les variations morphologiques (pluriel, conjugaison…etc) et étend la recherche aux mots synonymes ou de même racine étymologique. L'orthographe est prise en compte : l'usager peut écrire « chaise » ou « cheze ». Ce moteur discerne le singulier du pluriel.

Un important travail préparatoire a été fait sur les différentes bases. Entre autre, la création de dictionnaires. Les requêtes sont systématiquement étendues aux mots proches, en

utilisant des dictionnaires pour construire le lien entre les terminologies spécifiques et le langage courant.

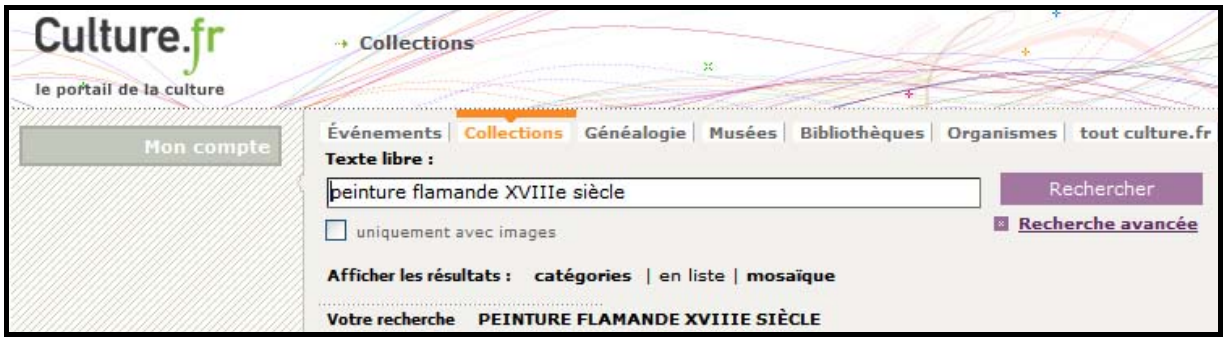
Comme chaque base utilise des formats documentaires différents, les champs sur lesquels il était indispensable que l'on puisse interroger ont été listés. Ainsi les quatre filtres fondés sur les métadonnées sélectionnées apparaissent : *qui, quand, catégories (type de document), où.*

Ainsi que les *termes associés* : le moteur calcule automatiquement les termes qui reviennent le plus souvent, en plein texte.

Portail Collections. Affichage des résultats. Critère de recherche : « femme »

⇒ Ces métadonnées non visibles pour l'internaute permettent de limiter le bruit via des pavés d'affinage de résultats (termes associés, qui, quand, où) avec des suggestions contextualisés.

A-t-on besoin d'un formulaire de recherches avec divers critères dont celui des « mots-clés » ? L'interrogation se fait en une seule fois et en langage naturel : « peinture flamande XVIIIe siècle ». Elle est par définition multi-critères. Il est inutile d'utiliser des opérateurs booléens.



Portail Collections. Encart de recherche, présent dans le bandeau supérieur du site.

2.3.2 Le "Laboratoire d'idées" d'Europeana

Le prototype actuel a cependant un « laboratoire d'idées » (Thoug Lab) qui propose une solution sur un sous-ensemble des contenus d'Europeana dans le domaine des musées.
http://www.europeana.eu/portal/thought-lab.html

Il utilise l'outil ClioPatria qui est un moteur de recherche sémantique. Pour ce sous-projet, les données fournies à Europeana sont celles du Rijksmuseum, du Musée du Louvre et du Rijksbureau voor Kunsthistorische Documentatie aux Pays-Bas. Ces métadonnées ont été utilisées dans leur format d'origine d'export vers Europeana (qui est requis en XML seulement et pas nécessairement en Dublin Core) et transformées en RDF. Toutes les informations de champs contrôlés ont été utilisées en association avec les dictionnaires correspondants.

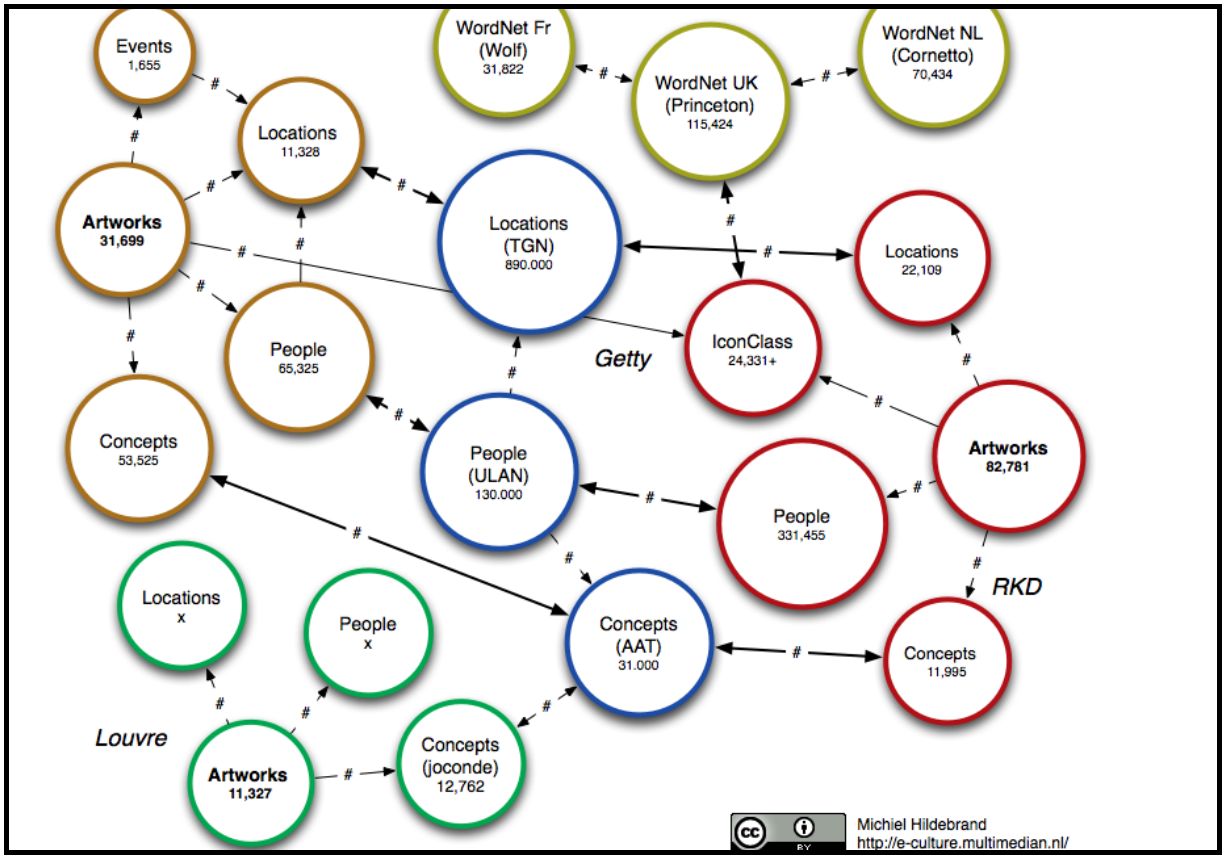
Pour le Louvre, par exemple, il y a eu contribution de deux bases : Joconde et Louvre Atlas (seules les descriptions d'objets en commun entre les deux bases ont été conservées).
Joconde contient des descriptions savantes et la liste des champs contrôlés, dont en particulier la description des images selon le thesaurus Garnier.

Louvre Atlas est l'application de visite virtuelle du Musée. Les métadonnées sont narratives et orientées tout public.

Une description unique a été créée pour chaque objet en faisant la somme des notices de chaque base.

Les thesaurus associés et leurs relations sémantiques sont décrits à http://eculture.cs.vu.nl/europeana/www/datacloud.html. Ils sont utilisés pour la traduction des concepts, des lieux et des noms propres afin de permettre la recherche multicritères des ressources hétérogènes (thesaurus et langue).

Les thesaurus utilisés sont : Joconde; IconClass; RKD Artists; le Getty: AAT pour les mots-clés, ULAN pour les noms d'artistes, TGN pour les noms géographiques; WordNet pour les termes en hollandais, anglais et français.



Laboratoire d'idées. Data cloud : Thesaurus associés et leurs relations sémantiques

Si je cherche le mot-clé **Mariage** dans le Thought Lab (laboratoire d'idées), le moteur me propose des résultats regroupés par proximité sémantique :

-œuvres évoquant le concept recherché (mot-clé est dans le titre ou le critère sujet)

-œuvres dont le mot-clé est dans le titre

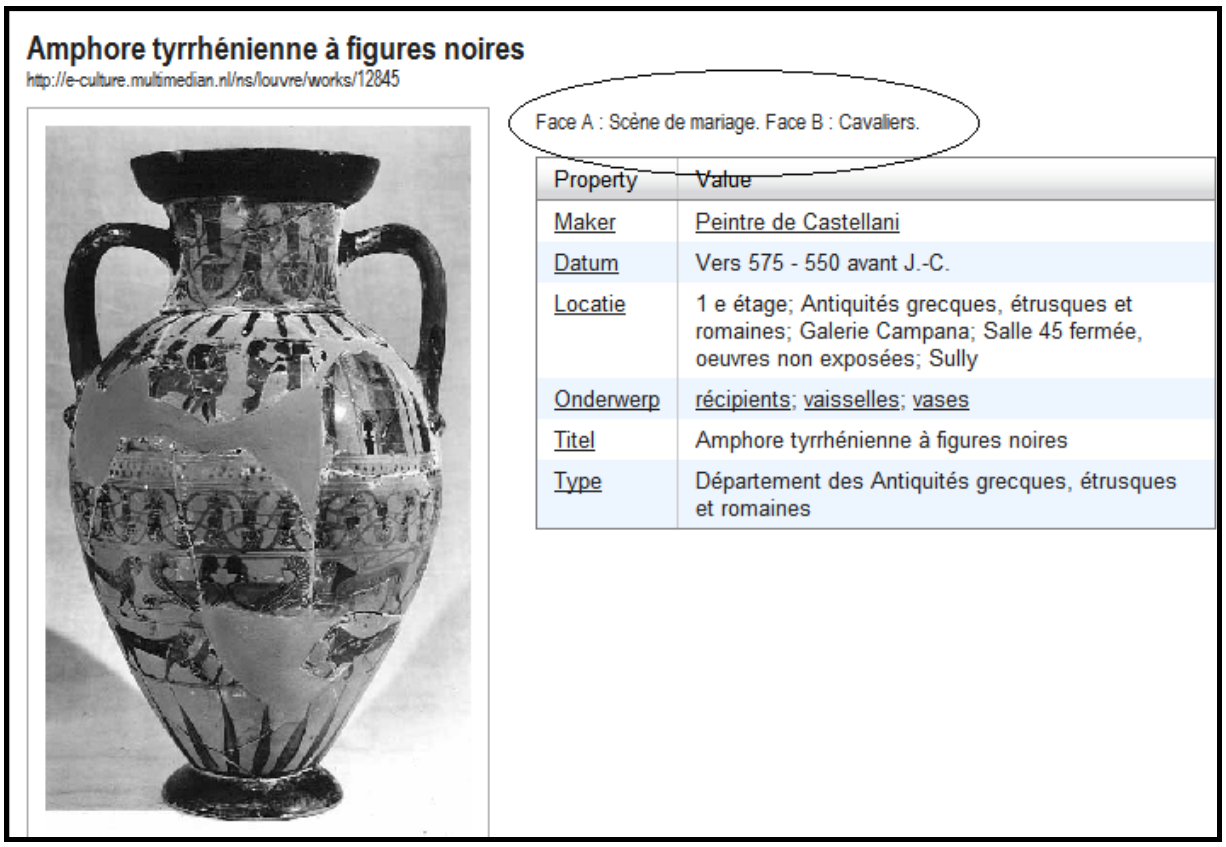
-oeuvres évoquant un élément plus spécifique : marriage ceremonies, public festivities at marriage of royal persons

-oeuvres ayant un lien avec le concept : danse, fête, couple. *Eté* et *Faux-pas*

-œuvres dont l'aspect suivant correspond : note

Les thesaurus et index sont en format SKOS.

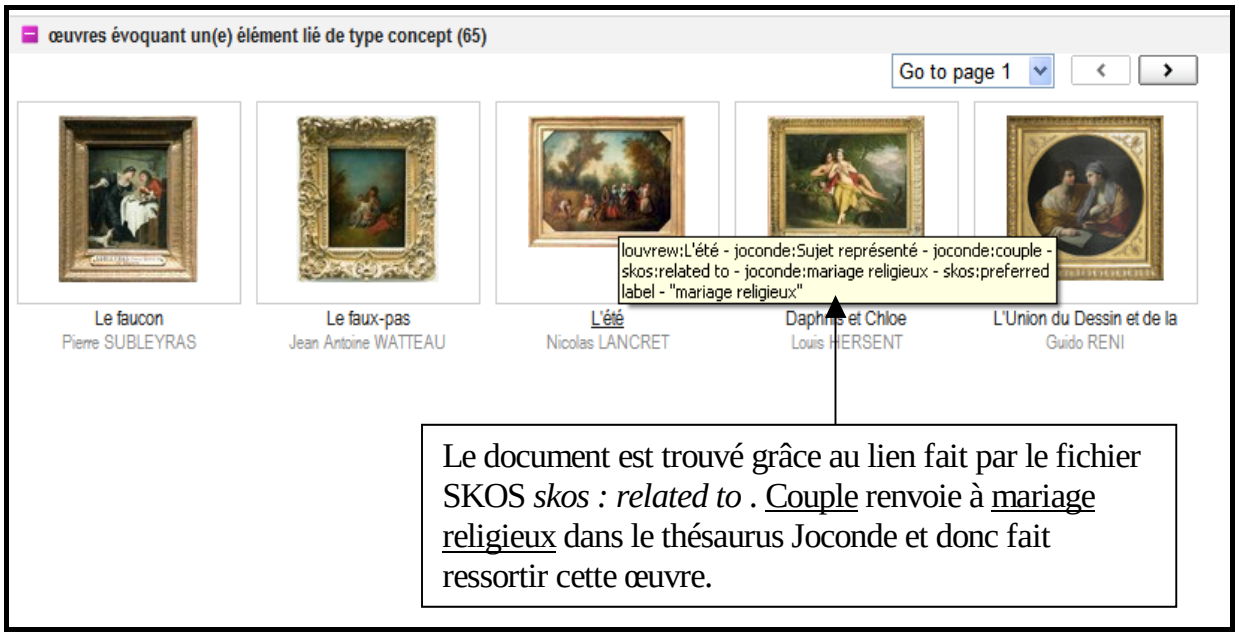
Œuvres dont l'aspect suivant correspond : note



Affichage de résultats. Laboratoire d'idées. Le terme mariage apparaît dans le champ « Note ».

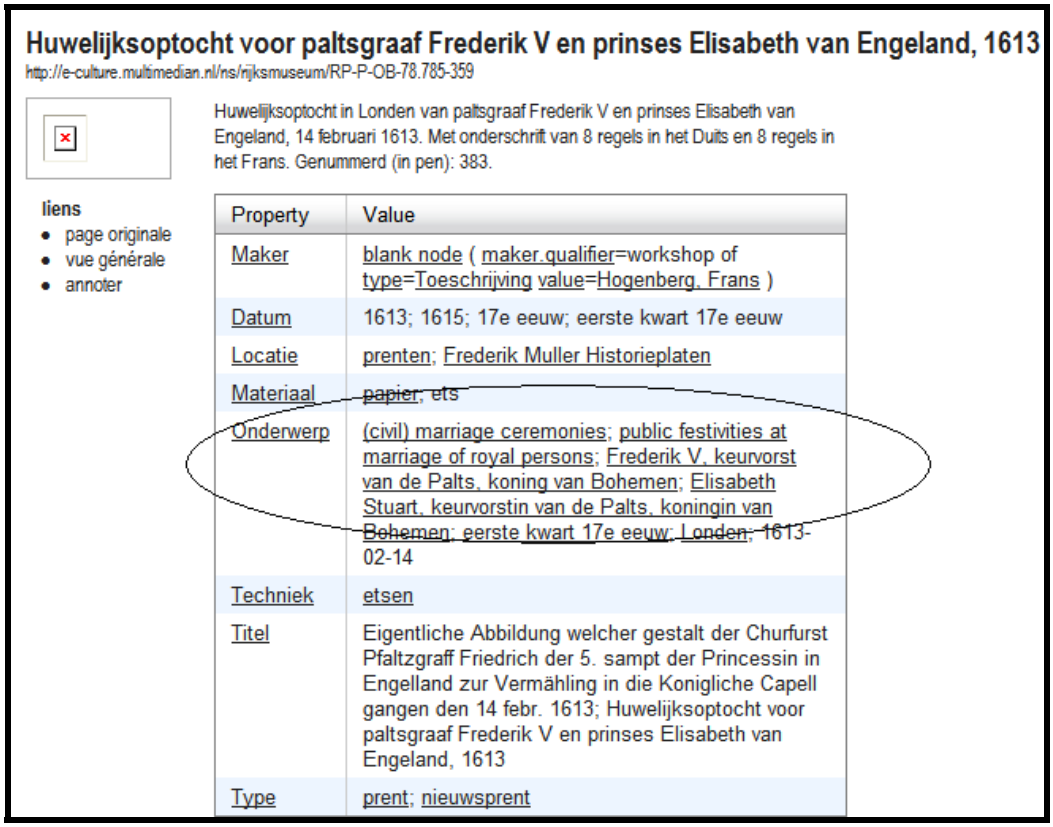
Œuvres évoquant un élément lié au concept « mariage ».

Le Faux-pas de Jean-Antoine Watteau et *L'Ete* de Nicolas Lancret. Le premier a comme mot-clé : couple, galanterie, scène, étreinte, peintures. *L'Ete* de Nicolas Lancret a comme mot-clé : couple, danse, fête etc..

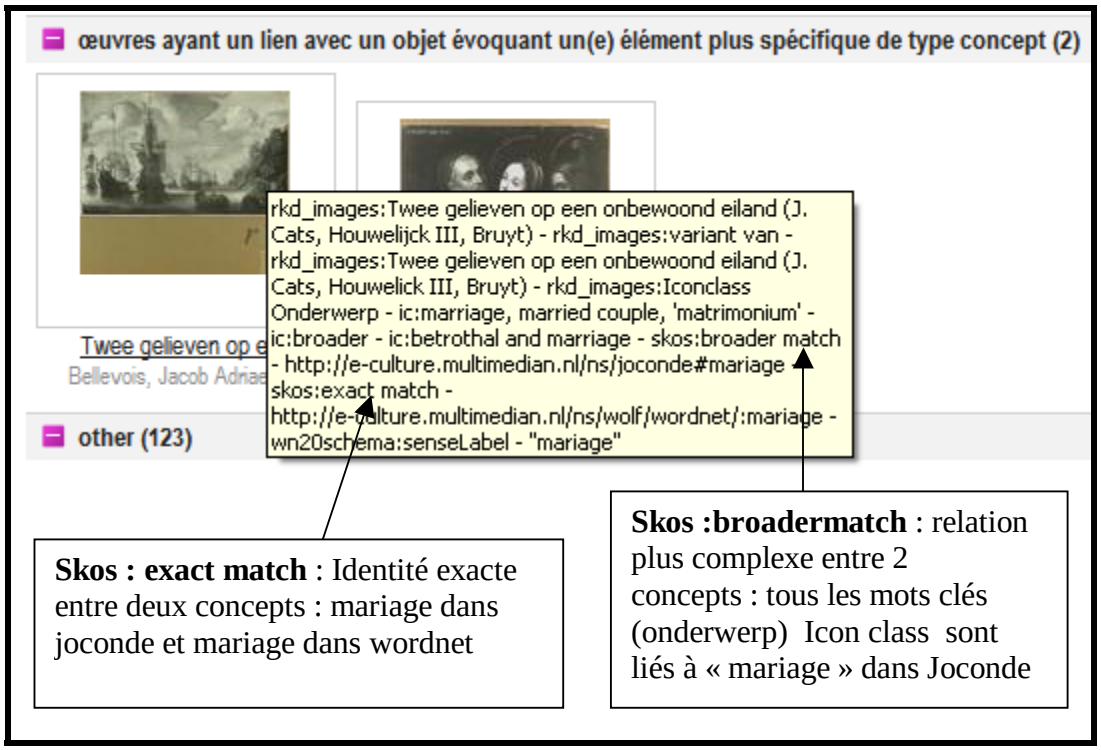


Affichage de résultats. Laboratoire d'idées. Alignement des vocabulaires contrôlés par le standard SKOS.

Œuvres évoquant un élément plus spécifique : (civil) marriage ceremonies ; public festivites at marriage of royal persons.



Notice d'œuvre. Mot-clé.



Affichage de résultats. Laboratoire d'idées. Alignement de terminologies par le fichier SKOS.

Mot-clé de cette notice : literatuur, zeegezicht, Kust, eiland, echtpaar

L'analyse de l'existant montre que chaque institution a adapté l'accès à ses contenus selon ses spécificités. L'idéal est d'avoir un thésaurus transversal mais lorsqu'il n'y a pas de langage d'indexation commun aux différentes applications, plusieurs solutions se dessinent:

-les bases ayant des champs communs sont regroupées : médiathèque, photothèque, phonothèque et sont indexées avec un langage documentaire commun. Les bases « archives » et « objets » sont administrées à part.
-si l'interface de recherche propose un accès simultané à des types de support aussi varié que des imprimés, des œuvres (œuvres d'art, objets mobiliers ou immobiliers), archives, il se peut qu'il n'y ait pas d'harmonisation de mots-clés.

Une troisième solution est l'apport des technologies dites sémantiques. Elles semblent être une solution pour aligner ou mettre en correspondance les vocabulaires (le laboratoire d'idées d'Europeana) ou faire du traitement automatique de la langue (analyse syntaxique, morphologique et sémantique) comme le moteur de recherche *Sinequa* du portail *Collections*.

Que peuvent apporter ces technologies pour remédier à la question de l'harmonisation des langages documentaires dans la recherche multi-sources ?

Troisième partie
Le web sémantique : une solution d'avenir ?

1 Contexte de l'étude : le MuCEM et la création de son portail documentaire

1.1 Le MuCEM : un projet scientifique et culturel

Le Musée national des Arts et Traditions populaires (MNATP) situé à Paris, près du Jardin d'acclimatation, créé en 1937 par Georges-Henri Rivière, présentait dans ses galeries une vision synthétique de la société traditionnelle française, rurale et régionale, du Moyen-Age aux années 50. En 2001, la chute durable de la fréquentation du MNATP a imposé de repenser l'institution. Il est alors décidé, sur proposition de la direction du musée, l'élargissement de son territoire géographique et l'extension de son propos. Au MNATP succède le Musée des civilisations de l'Europe et de la Méditerranée (MuCEM). C'est un musée de société à vocation de synthèse, par opposition aux musées thématiques ou dédiés à un territoire plus restreint, qui a pour mission de présenter au grand public comme aux scientifiques les grandes questions d'aujourd'hui en comparant et en interprétant les objets de la vie quotidienne dans l'espace euro-méditerranéen.

Le projet scientifique et culturel a été élaboré par le directeur du musée : Michel Colardelle et par son équipe, enrichi par un Comité scientifique composé de spécialistes nationaux et internationaux.

Le musée a élargi ses domaines d'investigation dans l'espace, passant de l'Europe à la Méditerranée, et dans le temps, depuis le Moyen-Age jusqu'au monde contemporain. Sa thématique a, elle aussi, évolué passant de la campagne à la ville, ainsi que ses moyens méthodologiques : l'ensemble des sciences sociales y seront représentées même si l'ethnologie restera la discipline centrale. Son objectif premier est de donner à chacun de nouveaux repères pour approfondir ses interprétations du présent [47, Colardelle].

Conçu dans le cadre administratif de l'Etablissement public d'aménagement *Euroméditerranée*, en charge du grand projet de développement national et européen de Marseille, ce nouvel établissement doit être pensé comme un musée international et un centre de culture vivante[35]. Cinq thèmes (EER[36]) devraient composer les expositions de référence qui seront présentées au public durant un premier cycle de quelques années : Figures du paradis, l'Eau, la Cité, le Chemin, Masculin-féminin. Ces expositions seront

[35] Situation et programme muséographique au 01.06.2009
[36] Expositions Evolutives de Références. Ces présentations sont appelées à évoluer.

ensuite transformées et porteront sur d'autres questions de société. Le musée fonctionnera comme un forum, un lieu de débats, où les expositions de référence (3 à 5 ans) et les expositions temporaires (5 à 6 mois) s'articuleront autour de grands sujets de société. Aux objets exposés sera associé un ensemble d'activités : programmation de films de cinéma et de spectacles, cycles de conférence et de rencontres, créations multimédias [45 ; 46, Colardelle].

Le MuCEM doit quitter Paris et s'installer à Marseille, à l'entrée du Vieux-Port, dans un bâtiment en dentelle de béton, construit par les architectes Rudy Riccotti et Roland Carta. Ce bâtiment sera relié par une passerelle au Fort Saint-Jean, monument historique, qui abritera le futur centre de ressources. Le MuCEM ouvrira ses portes en 2013. Le futur musée est intégré dans le schéma de réaménagement urbain *Euroméditerranée* et trouve sa place dans le projet d*'Union pour la méditerranée* du président Nicolas Sarkozy. En juillet 2007, Michel Sappin, le préfet des Bouches-du-Rhône, avait signé le permis de construire du musée, mais les travaux n'avaient pas pu démarrer. Grâce à l'élection de Marseille-Provence comme capitale européenne de la culture en 2013, le gouvernement a relancé le projet, inscrit également dans le plan de relance de l'économie française présenté début décembre 2008.

1.2 Le système d'information et de documentation : un portail documentaire à créer

Actuellement, le MuCEM continue officiellement à appliquer le même organigramme que celui de la fin des années 90, c'est pourquoi, il dispose toujours sur le papier d'un service des collections, d'une iconothèque, d'un centre de documentation, d'une bibliothèque, d'une phonothèque, d'un service historique (archives), d'un service audiovisuel et d'une photothèque. Le département « informatique et multimédia » (5 personnes), qui a pris en charge la coordination générale de la numérisation des ressources, le projet Anthroponet, la rénovation du portail du musée et l'édition multimédia destinée au public, est un service en préfiguration.

Au cours des dix dernières années, par manque de moyens financiers, chaque service a, soit utilisé, faute de mieux, des applications supportées par le DSI, mais techniquement dépassées comme Micromusée et les bases Mistral, soit des outils informatiques spécifiques (bases Filemaker, base 4D, Cadic intégrale, applications bureautiques) afin de répondre à ses propres besoins de catalogage et de numérisation et à la nature des objets et documents conservés.

Le projet de création d'un centre de ressources au MuCEM à Marseille implique une rénovation complète des systèmes d'information actuels et l'interopérabilité des futurs outils d'inventaire des collections et des fonds. Le but est d'offrir un service au public lui permettant d'accéder à des notices normalisées associées à des images des documents originaux, mais aussi à des enregistrements sonores ou audiovisuels. Il s'agira également de proposer à l'utilisateur un outil de recherche fédérée et croisée, en interrogeant simultanément plusieurs catalogues et inventaires, par exemple sur un sujet, une région, un auteur, un type d'objet et ses documents associés.

Le projet de service « informatique et multimédia » du MuCEM, qui a débuté en juin 2008, a souligné les grandes priorités et besoins pour faire progresser le chantier documentaire du centre de ressources. Ces priorités se mettent en place petit à petit [48, Dalbéra].

- disposer d'une assistance à maîtrise d'ouvrage (AMO) pour élaborer le schéma directeur informatique du futur musée et le cahier des charges pour rénover les outils documentaires, tout en prenant en compte les besoins de la période transitoire jusqu'au déménagement des collections à Marseille,
- terminer au plus vite la saisie informatisée des inventaires et catalogues de ressources encore tapuscrits ou manuscrits (*notamment pour les photographies et les enregistrements sonores*),
- accélérer le chantier des collections et engager la numérisation des dossiers d'œuvre,
- structurer en EAD[37] (*encoded archival description*), les instruments de recherche et inventaires du service historique,
- établir les priorités en matière de numérisation des documents primaires,
- étudier, avec le département des systèmes d'information du ministère, les solutions temporaires qui conduiront en quatre ans au futur service documentaire, intégré au portail actuel.

Les services actuels et la numérisation des ressources :

-La bibliothèque du MuCEM est l'héritière d'une partie des collections de l'Office de Documentation Folklorique créé en 1937 au Palais du Trocadéro. Le champ principalement couvert par la bibliothèque est celui de l'ethnologie française.

-Le service historique du MuCEM réunit plusieurs collections.

[37] Cette structuration des données, conçue pour les archives, a comme objectif la description de fonds et de leurs niveaux plutôt qu'une description de pièces

Les manuscrits[38] ; les photographies : 480 000 clichés argentiques et 140 000 cartes postales ; les fonds d'archives : les fonds d'archives publiques (histoire du musée), les fonds relatifs aux enquêtes ethnographiques, les fonds d'archives privées, voire familiales (fonds Arnold Van Gennep, fonds Georges Henri Rivière, fonds Marcel Maget).

-La photothèque, c'est-à-dire le fonds photographique et celui des cartes postales, fait partie du service historique, mais les fonds sont gérés indépendamment. L'essentiel des collections est constitué par des images prises au cours d'enquêtes ethnographiques réalisées depuis 1937, en France, par les chercheurs ainsi que par les photographies d'objets conservées au musée. Les collections intègrent également depuis peu des photographies d'enquêtes portant sur d'autres pays européens que la France ainsi que sur les Etats non-européens du pourtour méditerranéen.

-Le service audiovisuel conserve 700 films dont le plus ancien date de 1935. Ces films sont sur plusieurs supports : 16mm, 8mm, Super 8 et tous les formats vidéos. Actuellement 355 films sont sur support numérique.

Dans leur grande majorité, ces films ont été réalisés au cours d'enquêtes ethnographiques ou à l'occasion d'expositions présentées dans le Musée.

-L'iconothèque conserve un fonds d'estampes (environ 110 000 documents) qui comprend une des plus importantes collections françaises d'imagerie populaire. Elle possède également un fonds de dessins ethnographiques ainsi qu'une collection de peintures.

-Le service des collections est chargé du récolement général inscrit depuis 2003 dans le cadre légal et obligatoire de la loi sur les musées de France, et du traitement des collections jusqu'à leur conditionnement à travers le passage par une chaîne de traitement. Il s'occupe également des prêts et dépôts.

-La phonothèque : les premiers fonds sonores datent de 1943, mais les enregistrements les plus anciens remontent à 1939. La phonothèque a été créée par Claudie Marcel-Dubois, ethnomusicologue du CNRS. C'est à la fois un service de recherches qui a pour but de comprendre les phénomènes musicaux et de réaliser les inventaires des collections d'instruments de musique et un service d'archives qui doit conserver et mettre à disposition du public ses enregistrements sonores.

[38] On entend par ce terme tapuscrits, photos et manuscrits à proprement parler.

Récapitulatif : 8 bases de données, 4 langages contrôlés, deux listes d'indexation et des formats documentaires différents selon chaque base

Service	**logiciel**	**Format des notices**	**Indexation**	**Base de données**
Bibliothèque : Ouvrages et périodiques	Cadic-Intégrale	Unimarc	Rameau	Catalogue web consultable sur le réseau local
Objets : service des collections	Micromusée	Format DMF	Thesaurus « Système descriptif des objets domestiques français » et Liste DMF	Une petite partie est exportée dans Joconde Internet
Phonothèque	File Maker	Format maison	Pas d'indexation matière	Consultable sur place
Vidéothèque	File Maker	Format maison	Ethnophoto Thesaurus Garnier Liste maison	Consultable sur place
Service historique	Word Arkéïa	Normes Archivistiques EAD		Consultable sur place
Photothèque	Mistral-Editor	Format maison compatible Joconde	Thesaurus Garnier	Phocem Internet
Cartes postales : service historique		Format maison Compatible Joconde	Thesaurus garnier et ethnophoto	Carpo et Joconde Internet
Iconothèque : Service des collections	4D	Format DMF compatible Joconde	Thesaurus Garnier	Une petite partie est exportée dans Joconde Internet

2 Réalisation d'une terminologie d'indexation pour le corpus masculin-féminin

2.1 Les enquêtes-collectes sur le mariage et les rites de passage

Dans son programme muséographique, le MuCEM a choisi de traiter le thème *Masculin-Féminin : la construction du genre*. Dans toutes les sociétés, la différence biologique entre les sexes est utilisée pour construire des catégories sociales qui renvoient à l'assignation de statuts et de rôles pour les hommes et les femmes [52, Héritier]. Ces statuts sont inscrits dans des ensembles de pratiques et de représentations ; ils irriguent les sociétés au point que le marquage masculin ou féminin est omniprésent, tant dans les actes les plus ordinaires de la vie quotidienne que dans les moments où les sociétés se mettent en scène pour mieux affirmer leurs spécificités.

En 2005 est lancée la campagne *Masculin-Féminin : la construction du genre*. Dans le cadre de cette campagne, des enquêtes-collectes sur le mariage et sur des couples vivant dans l'espace euro-méditerranéen se mettent en place. Denis Chevallier, directeur de l'antenne du MuCEM à Marseille, en est le responsable scientifique. Il est assisté de Marina Zveguinzoff pour la coordination. Une équipe de recherches s'est mise en place ainsi qu'un comité scientifique. De nombreuses réunions ont lieu pour définir les objectifs, les méthodes de recherches (choix des terrains et élaboration des outils d'enquêtes), la présentation des résultats (choix des objets, restitution des résultats sous forme de dossier numérique)[39]. 14 enquêtes eurent lieu en 2005 et 2006.

Au-delà du mariage, de nombreux rites participent à la construction de l'identité sexuée de la personne, l'identité ne dépendant pas seulement du donné biologique, mais se construisant aussi socialement. De la naissance à la mort, l'identité de la personne féminine ou masculine d'un individu est façonnée par l'apprentissage de pratiques, de gestes, de postures en relation avec un univers matériel spécifique (jeux, outils, vêtement, mobilier…) et par un ensemble de rites. C'est pourquoi la campagne *Masculin-Féminin* a pris en compte tous les rites de passage participant à la construction de l'identité [51, Chevallier].

[39] Rapports d'étapes 2005 et 2006

L'objectif final de cette enquête est de faire une exposition qui comprendra deux sections « devenir homme et femme aujourd'hui : apprentissage et passage » et « histoires de genre ». Plus d'une centaine d'objets ont été sélectionnés pour y figurer[40], certains appartenant à la campagne-collecte *Masculin-Féminin*, d'autres aux collections du MuCEM.

Cependant, dans un premier temps, l'objectif est de faire une publication multimédia. En effet, une **collection multimédia** [41] a été créée en 2005. Elle traduit la volonté de valoriser les collections, de moderniser l'image du musée tout en favorisant la diffusion des résultats des recherches ethnologiques et en accompagnant la politique d'exposition de préfiguration. Le site *Masculin-Féminin, histoire de couple et construction du genre* fait partie de cette collection et a été mis en ligne en 2009[42].

2.2 Le site Masculin-Féminin, histoire de couple et construction du genre : public cible et accès à l'information

Cette publication multimédia doit amener l'internaute à découvrir la question des modalités de la construction du genre à partir de situations observées aujourd'hui et des rites de passage qui persistent dans nos sociétés contemporaines. Ce site est conçu comme un lieu de diffusion de données ethnologiques dont le public cible est un public non spécialiste. Il fait connaître au grand public les résultats des recherches sur ce thème faites depuis 2005, sous la responsabilité scientifique de Denis Chevallier. Les usagers ont accès à la variété des supports : photos, objets, enregistrements sonores, vidéos collectés au cours des enquêtes de terrain, et à d'autres sources d'informations : sites web, bibliographie, filmographie.

Ce site a été financé par la MRT[43] et la campagne de recherches *Masculin-Féminin* par la RMN[44]. Il est administré de manière dynamique par un système de gestion de contenu (CMS) open source eZ publish. Le pôle multimédia est chargé de la conception de ce site. Ce service est sous la responsabilité de Jean-Pierre Dalbéra et se compose de quatre personnes : Mohan Danabalou, responsable informatique du MuCEM, Yannick Vernet, chef de projet multimédia et Filippo Vancini, concepteur multimédia.

---

[40] Programme muséographique, sous la direction de Michel Colardelle, daté du 20 juin 2007.
[41] http://www.ethnologie.culture.fr
[42] <http://www.femininmasculin.culture.fr>
[43] Mission de la recherche et de la technologie
[44] Réunion des musées nationaux

Le site Masculin-Féminin, histoires de couple et construction du genre, est structuré de la façon suivante :

Les couples : on décrit l'histoire de chaque protagoniste et leur rencontre. Quatorze couples et neuf pays ont été sélectionnés pour figurer sur le site.

Les rites de passage : cette rubrique se compose de deux parties : « rites et genre » et « autres rites ». On explique les rites qui, dès la naissance, participent à la construction de genre et à l'élaboration de l'identité sociale : La naissance, l'attribution du nom, la circoncision, la coupe de cheveux, la majorité religieuse, le service militaire, la fin de vie de célibataire et le mariage. D'autres rites façonnent notre identité : le 7ème jour, le baptême, le rentrée des classes, la communion catholique ou laïque, la fin de cycle scolaire...

Les différents types d'unions : mariage, PACS, concubinage.

L'architecture de ce site a été validée par les chercheurs au cours de deux réunions qui ont eu lieu en 2007. Concernant la question des droits, il a été convenu que chaque image fixe ou animée sera protégée par un copyright. Les chercheurs et les couples ont autorisé la diffusion de l'information quelque soit le support : objet, vidéo, son, texte, photo, à condition de mentionner l'auteur de l'œuvre.

2.3 Elaboration d'un lexique transversal

2.3.1 La bibliothèque numérique OMEKA

Omeka est une bibliothèque numérique « open source », développée par le CHNM[45] et qui fonctionne sous Linux, Apache, MySQL 5.0+, PhP 5.2+. Cet outil est destiné à mettre en ligne des corpus multimédias pour les institutions culturelles et scientifiques. Il sera prochainement doté d'une interface OAI-PMH et pourra être moissonné selon ce protocole. L'application a été installée sur le serveur du MuCEM. Cette bibliothèque numérique a été choisie pour tester l'indexation d'images qui ne relevaient pas de la base Phocem. Ce travail a permis d'expérimenter la méthode de catalogage de l'ensemble des documents multimédias au format Dublin Core et la correspondance des champs des diverses bases dans ce même format.

[45] Centre pour l'histoire et les nouveaux médias

2.3.2 Catalogage des documents: difficultés rencontrées

2.3.2.1 Correspondance des champs

Dans Omeka, les métadonnées sont indexées au format Dublin Core (simple). Il reste donc à faire correspondre les différents champs de divers formats de chaque base en un format Dublin Core.

Deux tables des correspondances ont été faites en 2007. L'une par Julia Bontempi, stagiaire au MuCEM et l'autre par Geneviève Deblock, conservatrice de la bibliothèque [44, Bontempi]. Cataloguer les différentes ressources en Dublin Core a présenté quelques difficultés. Mais les ajustements et les solutions étaient faciles à trouver. Par exemple, les champs de la notice de la base photo ne correspondent pas nécessairement au format Dublin Core.

Le champ « Légende », en base Phocem, correspond au champ « Description », en Dublin Core. Mais le champ « Titre » en Dublin Core n'existe pas dans la base Phocem. Or, il est nécessaire de le remplir. En effet, ce champ est un des critères de recherches du moteur de OMEKA, et ce critère est commun à tous les documents. Si l'utilisateur interroge par « mots du titre », aucune image fixe n'apparaîtra si le champ n'est pas rempli. La solution est de répéter une partie du champ « description » si nécessaire ou d'inscrire un titre d'ensemble.

2.3.2.2 Les règles d'écriture

Des règles syntaxiques sont à définir pour certains champs comme « creator » et « contributeur » et de manière générale, pour tous les noms de personne. On s'est basé sur la notice d'autorité « nom de personne » de la BNF : Nom, Prénom. Pour les champs indiquant une date, Omeka impose des règles d'écriture :

-YYYY MM DD (année, mois, jour) pour le champ « date »

-from YYYY MM DD to YYYY MM DD pour le champ «temporal coverage»

2.3.2.3 Choisir un langage d'indexation commun : "sujet" et « lieu »

Pour indexer des photos et tout ce qui concerne l'iconographie, le MuCEM utilise le thésaurus Garnier. Or, il n'est pas adapté au public visé pour le site internet. Certains termes ne sont pas dans le thésaurus comme « mariés » ou « mariée ». Isabelle Gui, responsable de la photothèque, a proposé de faire une demande d'ajout de descripteur au thésaurus Garnier. Si on a besoin de ces termes, il est possible de les intégrer dans Garnier. Pourtant cette solution ne résout pas le problème de la syntaxe de Garnier qui n'est pas adapté au grand public.

Exemple du champ « mot-clé » dans la base de donnée Phocem :

- portrait collectif (mariage, époux, épouse, debout)

-scène (mariage, homme, femme, enfant, prêtre, statue : sainte) ; représentation d'objet (peinture, tableau, 16e siècle)

De nombreuses institutions, comme la BPI et le Musée du quai Branly, utilisent le thésaurus Rameau pour indexer ce genre de support. Photos, sites web, vidéos sont catalogués avec Rameau. De ce fait, on accède via la recherche fédérée, à différents contenus : ouvrages et imprimés divers, iconographie, documents sonores et visuels. Mais le langage d'indexation Rameau, certes plus simple dans sa syntaxe, n'est pas vraiment adapté pour indexer des images fixes ou animées avec précision. Il permet néanmoins une harmonisation des mots-clés.

La liste des normes utilisées pour le champ « sujet » en Dublin Core est la Dewey, la classification de la bibliothèque du Congrès, l'Universal Classification (une Dewey élaborée). Mais ces langages d'indexation sont trop généraux pour notre sujet et inadaptés.

⇒ Au final, il a été décidé de constituer une liste de mots-clés propre au corpus Masculin-Féminin à partir du thésaurus Garnier, du *Système descriptif des objets domestiques* et du vocabulaire collecté dans les ouvrages sur la construction du genre[46]. Cette liste de termes a quelques relations de synonymie, mais l'on sait que le moteur de recherche Omeka et du site internet *Masculin-Féminin* n'ont pas cette fonctionnalité. Les relations sont donc minimalistes.

Pour le champ « lieu », on simplifie aussi l'indexation en utilisant le thésaurus Getty avec comme séparateur [,]. On suppose que la plupart des utilisateurs seront français, on inscrit le terme en anglais et en français.

Ce travail sur Omeka a permis de tester la méthode d'indexation au format Dublin Core et la correspondance des champs selon le type de documents. Les enseignements tirés de cette expérimentation ont servi à choisir les descripteurs des ressources figurant dans le site Masculin-Féminin développé sous la plate-forme eZpublish.

[46] cf. Bibliographie

2.4 Elaboration d'un lexique transversal avec des outils linguistiques automatiques

Cette création de terminologie s'inscrit dans le projet DAFOE dont la société Mondeca est un des partenaires.

2.4.1 Le projet DAFOE

L'objectif du projet DAFOE est de proposer une méthode complète associée à une plateforme technique pour concevoir des ontologies, de la modélisation à partir du domaine à leur évolution en passant par leur formalisation et exploitation. Le projet prend en charge la modélisation sémantique des concepts ontologiques.

La plateforme technique DAFOE est un ensemble d'outils dont un éditeur d'ontologies qui prend en charge toute la question de la sémantique de ces ontologies afin d'obtenir une ontologie formalisée qui pourra être traitée dans un éditeur d'ontologie respectant les standards des langages d'ontologies du W3C (OWL) [43, DAFOE].

L'évaluation finale de cette plateforme se fera via des applications pour lesquelles une ontologie est nécessaire et ce à travers sa dimension sémantique et donc son interaction avec l'utilisateur. Les applications développées correspondent à des tâches d'indexation de documents, puis de recherche d'informations à leur sujet. Ces applications seront mises en oeuvre dans trois domaines 1) l'aide au codage médical, 2) l'indexation patrimoniale et 3) l'indexation d'images satellitaires.

Les acteurs du projet sont : Jean Charlet (Institut national de la santé et de la recherche médicale, INSERM), ENST/GET (Paris), IRIT (Institut de Recherche en Informatique de Toulouse), LIPN (Laboratoire d'Informatique de l'Université Paris-Nord), LISI (Laboratoire d'informatique scientifique et industrielle, Poitier), Mondeca (Paris), Supelec (Saclay), UTC (Université de Technologie de Compiègne).
Il est financé par l'ANR (Agence nationale de la Recherche) et labellisé Cap Digital.

La plate-forme est en cours d'élaboration. Elle doit être opérationnelle en 2010 mais un premier protocole pourra être testé en novembre.

2.4.2 Le MuceM et MONDECA

Créé en 2000, Mondeca est un éditeur de logiciel spécialisé dans les technologies du web sémantique. Cette société a pour mission de fournir des logiciels et des services pour aider les entreprises à organiser, valoriser et publier leurs contenus.

Son outil est ITM (Intelligence Terminologie Management). ITM est un espace de travail qui sert à modéliser les ontologies. C'est aussi un espace de gestion de thésaurus, taxonomies, terminologies et bases de connaissances multilingues. Il stocke et gère les référentiels métiers [36, Mondeca].

Pour exploiter les bases de connaissances et valoriser leur contenu, Mondeca propose la mise en place de portail sémantique. Un portail sémantique est un site internet qui offre une porte d'entrée unique sur des ressources et des services centrés sur une base de connaissances. Par exemple, le site *Nièvre en Bourgogne*[47] est un portail sémantique permettant à l'utilisateur de rechercher et naviguer dans un espace informationnel en exploitant la sémantique et la base de connaissances multilingue élaborées en amont.

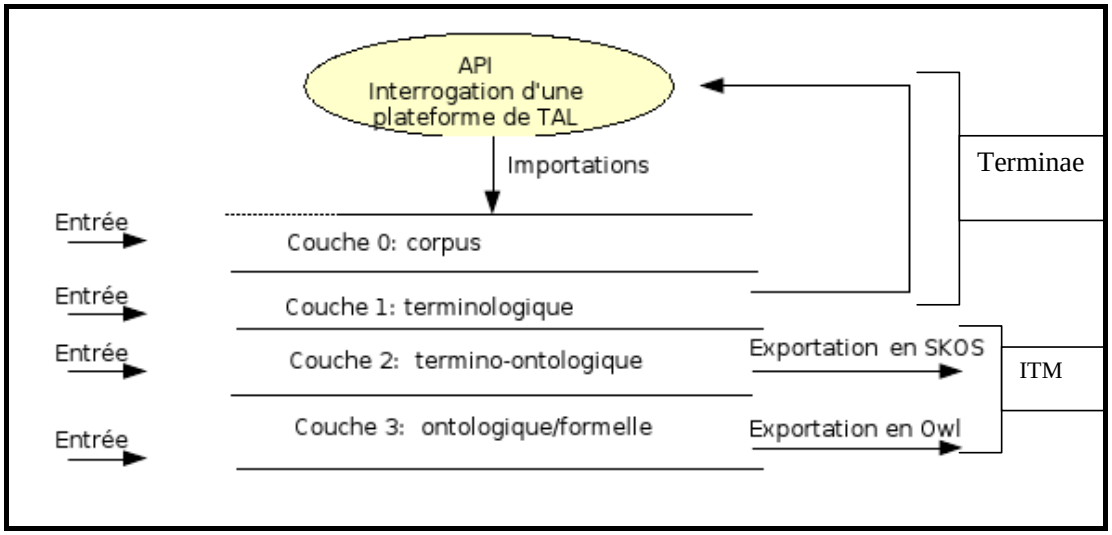
La société Mondeca travaille avec quatre catégories de clients appartenant aux secteurs suivants : presse, médias et droit ; industrie (Sncf, EDF, EADS…) ; santé ; tourisme (Nièvre, région de Bourgogne, projet Strabon[48]). Pour élaborer une ontologie, Mondeca a besoin de documents textuels qui seront analysés et croisés avec des thesaurii existants et des listes terminologiques.

Le projet DAFOE est l'occasion de créer une terminologie avec des outils de traitement automatique de la langue. L'évaluation des outils DAFOE est effectuée, dans le domaine de l'indexation culturelle et patrimoniale, avec le MuCEM. Le corpus choisi est *Masculin-Féminin : la construction du genre*, mais le projet peut s'étendre à d'autres vocabulaires d'indexation et couvrir par la suite l'ensemble des thématiques du musée. Un des buts futurs est d'enrichir le vocabulaire d'indexation actuel de façon semi-automatique.

---

[47] http://www.nievre-tourisme.com/

[48] http://strabon.tech.fr/spip.php?article8. Le projet STRABON doit créer une plate-forme multilingue et multimédia, accessible par Internet, dédiée au patrimoine culturel et aux activités touristiques des pays méditerranéens. Cette plate-forme offrira à chaque pays un



Cahier des charges scientifique et technique de la plateforme DaFOE - Chapitre Modèle de données. Terminae et ITM sont des outils Mondeca utilisés pour créer une ontologie.

2.4.3 Méthode d'élaboration automatique

Pour construire une terminologie, le processus de mise en œuvre est le suivant :

1. Corpus : choix du thème, identification des textes et des terminologies

2. Analyse automatique linguistique via l'outil « Tagger Tree » qui extrait les termes les plus pertinents.

3. Validation, définition, organisation et hiérarchisation des termes par les experts du domaine afin d'assurer la cohérence terminologique

4. Fusion entre la terminologie d'indexation issue du corpus et d'autres terminologies : Thésaurus Garnier, Terminologie Joconde des sujets représentés, Terminologie Joconde des domaines, Terminologie Joconde des dénominations, Structure géographique issue de GéoNames.

5. Nouvelle terminologie d'indexation, unifiée et mise à jour

La plate-forme DAFOE n'étant pas prête, les outils utilisés sont Tree Tagger Tree, Yatea et Terminae.

---

dispositif permettant de rendre interopérables ses propres systèmes d'information déjà en ligne et de construire de nouvelles bases de connaissances pour la culture et le tourisme.

2.4.3.1 Phase 1 : choix du corpus et structuration des données dans un fichier excel

Pour construire une terminologie, les documents suivants ont été sélectionnés :

-Textes du site internet *Masculin-Féminin : histoire de couples et construction du genre*

-Textes du rapport : La construction du genre en Europe et en Méditerranée : bilan de la campagne-mariage 2005-2008. 103 p.

-Articles de sociologues et anthropologues publiés dans la revue *Terrain* (consultables en ligne).

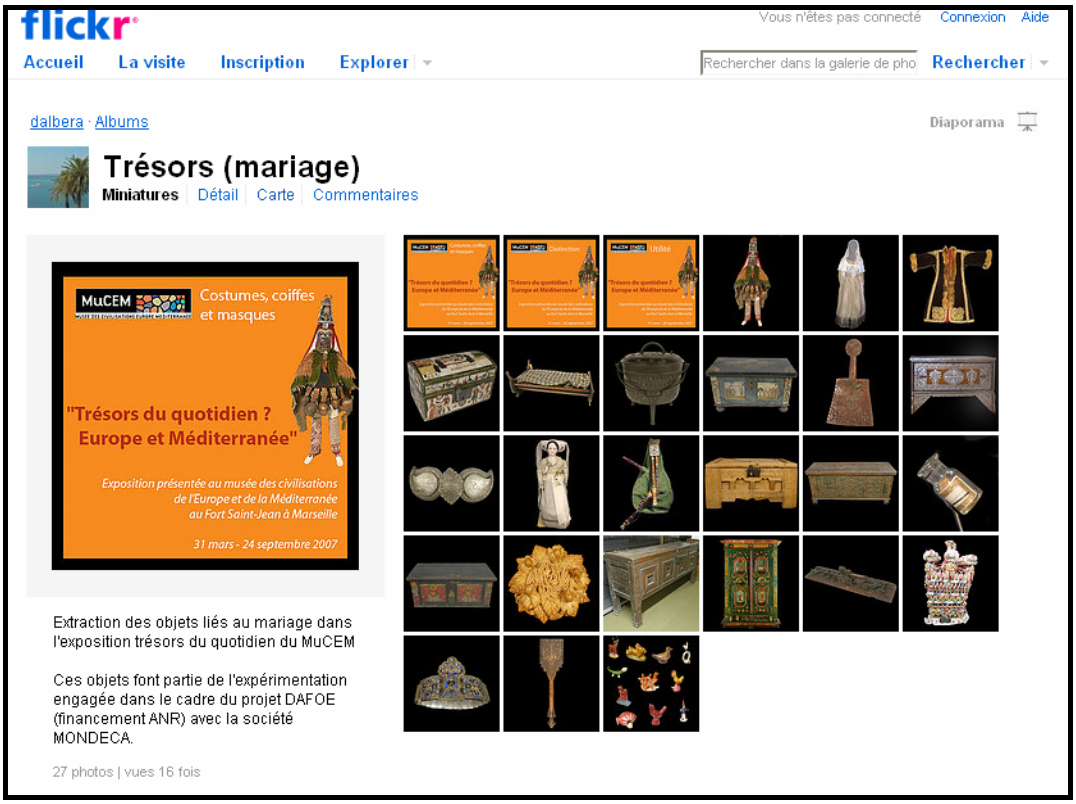
Boukhobza Noria. Dénouer les noces. Terrain [En ligne]. 2001, n°36, mis en ligne le 08 mars 2007.<http://terrain.revues.org/index1180.html> DOI : en cours d'attribution.

Flanquart Hervé. Un désert matrimonial. Terrain [En ligne]. 1999, n°33, mis en ligne le 09 mars 2007. <http://terrain.revues.org/index2710.html> DOI : en cours d'attribution

Hérault Laurence. La cheville et le brandon. Terrain [En ligne]. 1987, n°8, mis en ligne le 19 juillet 2007. <http://terrain.revues.org/index3152.html>DOI : en cours d'attribution

Nicolas Maud. Ce que « danser » veut dire. Terrain [En ligne]. Septembre 2000, n°35, mis en ligne le 08 mars 2007. <http://terrain.revues.org/index1065.html> <Consulté le 11 juillet 2009>

- Données des notices d'objets suivants :
    - objets de l'exposition « Trésors du quotidien » ayant pour thème le mariage et la construction du genre :
      <http://www.flickr.com/photos/dalbera/sets/72157621123444118/>



Flickr. Objets liés au mariage dans l'exposition Trésors du quotidien du MuCEM.

- objets de la campagne mariage : [Rapport sur le voile] La construction du genre en Europe et en Méditerranée : bilan de la campagne-mariage 2005-2008. 98 p.

Certains textes s'adressent à un public averti, d'autres à un public non spécialiste.

Toutes ces données doivent être en format XML afin d'être analysées automatiquement par l'outil linguistique Tagger Tree[49].

Pour cela, deux fichiers excel sont créés dans lesquels on structure l'information selon les catégories du site (page d'accueil, histoire de couples, cérémonies et sociabilités, cultures et tradition) et un format documentaire de 12 champs (Type de documents ; Titre ; Période ; Pays, Région, Ville ; Matériau 1 ; Matériau 2 ; Matériau 3 ; Description ; Origine ; Catégorie).

2.4.3.2 Phase 2 : Traitement automatique de la langue avec trois outils : Tree Tagger, Yatea et Terminae.

Une fois cette première phase terminée, le traitement automatique de la langue (TAL) est mis en œuvre en segmentant les mots et les phrases des textes afin de passer les données dans l'outil Tree tagger. Le programme python est utilisé pour cette segmentation. Il sera finalement abandonné car il pose des problèmes d'encodage.

[49] http://taln09.blogspot.com/search/label/Morpho-syntaxe

2.4.3.2.1 Analyse morphologique avec l'outil Tree Tagger

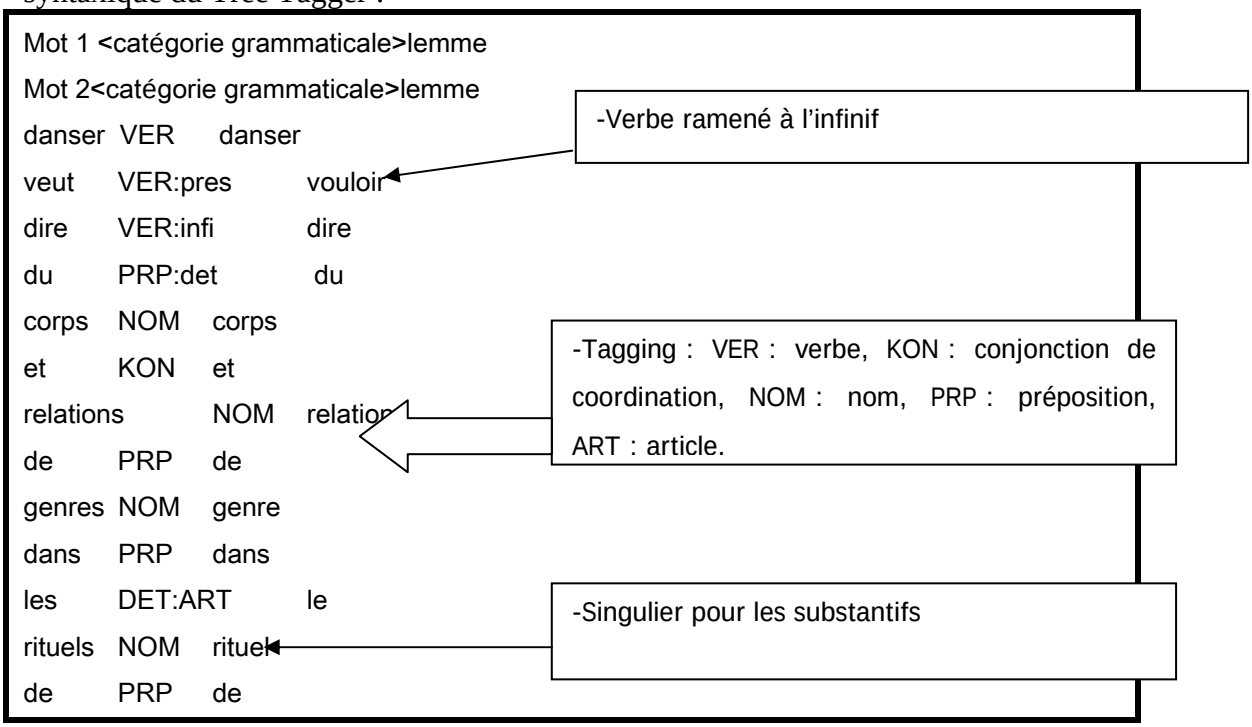
Une fois les mots identifiés par segmentation, l'analyse morphologique permet de reconnaître les formes des mots (tagging), c'est-à-dire d'identifier la catégorie d'appartenance des mots : nom, verbe, adjectif, adverbe etc…[41, Raïs].

Les termes identifiés font l'objet d'une lemmatisation : identification de la racine ou forme canonique d'un terme dans un dictionnaire pour pouvoir traiter les différentes variantes possibles. Le lemme correspond à l'infinitif pour les verbes, le singulier pour les substantifs, le masculin-singulier pour les adjectifs.

Une fois que le lemme est identifié, il est possible d'accéder à des dictionnaires et de traiter
-les formes fléchies : singulier/pluriel, masculin/féminin
-les formes dérivationnelles : termes construits autour d'une même racine (infinitif pour les verbes et singulier/masculin pour les substantifs).
Exemple :
Reconnaissance du pluriel : cheval/ chevaux
Reconnaissance des formes verbales ramenées à l'infinitif : veut/vouloir

A la fin de la lemmatisation, le texte est découpé en lemmes avec indication de leur nature grammaticale (nom, verbe, adjectif, adverbe…)

Extrait de l'article de Maud Nicolas (Revue *Terrain*) après traitement morpho-syntaxique du Tree Tagger :



mariage	NOM	mariage
Tunis	NAM	Tunis
Maud	NAM	Maud
Nicolas	NAM	Nicolas

2.4.3.2.2 Extraction des termes avec l'outil Yatea

Yatea est un outil d'extraction de termes développé au LIPN (Laboratoire d'informatique de Paris-Nord) par Sophie Aubin et Thierry Hamon.

A partir du fichier Tree Tagger, Yatea sort un fichier XML de termes qui vont servir à construire la terminologie. Cet outil examine les lemmes, les classe, enlève les mots vides (à, le, ils...), compte le nombre d'occurrence et repère l'emplacement des termes dans le texte.

Extrait du texte une fois traité par Yatea :

```
<TERM_CANDIDATE>
      <ID>term1850</ID>
      <FORM>rituels</FORM>
      <LEMMA>rituel</LEMMA>
      <MORPHOSYNTACTIC_FEATURES>
          <SYNTACTIC_CATEGORY>NOM</SYNTACTIC_CATEGORY>
      </MORPHOSYNTACTIC_FEATURES>
      <HEAD>term1850</HEAD>
      <NUMBER_OCCURRENCES>1</NUMBER_OCCURRENCES>
      <LIST_OCCURRENCES>
      <OCCURRENCE>
        <ID>occ3158</ID>
        <MNP>0</MNP>
        <DOC>0</DOC>
        <SENTENCE>0</SENTENCE>
        <START_POSITION>89</START_POSITION>
        <END_POSITION>96</END_POSITION>
        </OCCURRENCE>
      </LIST_OCCURRENCES>
      <TERM_CONFIDENCE>0.5</TERM_CONFIDENCE>
      <LOG_INFORMATION>YaTeA</LOG_INFORMATION>
    </TERM_CANDIDATE>
```

2.4.3.2.3 L'outil Terminae : sélection et validation des termes extraits par Yatea

Une fiche terminologique est créée. Elle regroupe toutes les caractéristiques d'un terme et les différents sens de ce terme dans le corpus. Chaque sens est décrit par un concept dit terminologique. Dans une fiche terminologique, il y a :

-des rubriques lexicales qui contiennent les caractéristiques lexicales d'un terme. Ces rubriques peuvent être ajoutées ou supprimées par l'utilisateur.

-les noms des concepts terminologiques. A chaque concept terminologique est associé un ensemble d'occurrences qui caractérise un sens d'un terme, un nom qui sert d'identifiant

dans l'ontologie associée, une définition en langage naturel saisi par l'utilisateur du système, des synonymes et des "voir aussi".

Il s'agit de regrouper, limiter, classer le vocabulaire afin qu'il soit régi par des relations de hiérarchie, d'équivalence ou de parenté entre les termes (descripteur générique, descripteur spécifique, descripteur associé). Les modalités d'emploi des descripteurs peuvent être brièvement expliqués dans des notes d'application (ou notes d'usage).

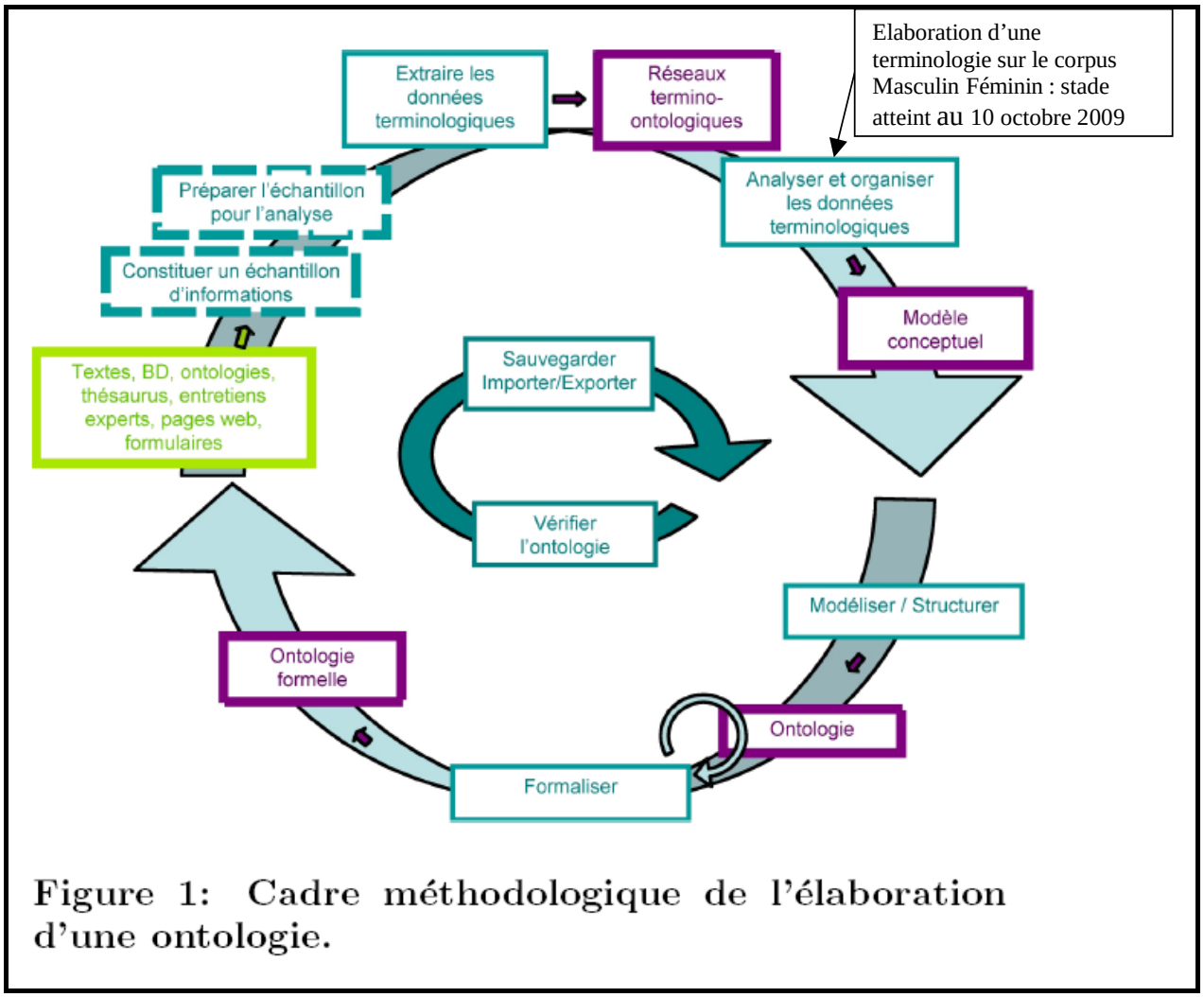


Figure 1: Cadre méthodologique de l'élaboration d'une ontologie.

DAFOE: A Multimodel and Multimethod Platform for Building Domain Ontologie - Jean Charlet, Sylvie Szulman, Guy Pierra [36, Mondeca].

Ce schéma présente les différentes étapes de création d'une ontologie. Actuellement, le projet d'élaboration d'une terminologie à partir du corpus Masculin-Féminin est au stade « Analyser et organiser les données terminologiques ». Cette étape est réalisée avec l'outil Terminae mais n'est pas automatique.

3 Préconisations pour le MuCEM

Dans le cadre de l'ouverture du MuCEM à Marseille, l'élaboration d'un schéma directeur des systèmes d'information va être menée.
Si une fonctionnalité de recherche fédérée est mise en place dans le futur portail documentaire, on pourrait suggérer, étant donné l'hétérogénéité des collections, de regrouper les huit bases distinctes en trois catégories et de faire correspondre certains champs. On pourrait avoir :
-une base *Médiathèque* qui fusionnerait les bases bibliothèque, photothèque, phonothèque et filmothèque. La norme Unimarc serait utilisée pour ces notices quelque soit le type de support.
-une base qui fusionnerait la base *objet* et la base *iconothèque*. Les notices seraient au format propre à la DMF[50] et au ministère de la culture.
-une base *Archives* en EAD/XML

La mise en place d'un moteur de recherche fédérée multi-bases/ multi-sources soulève un certain nombre de problèmes techniques. Mais la question du choix des protocoles de communication trouve un solution. On peut opter pour le principe du moissonnage ou celui des connecteurs.

Reste la question de l'harmonisation des langages documentaires du MuCEM.

3.1 Les différents scénarii

3.1.1 Solution « Musée du quai Branly »

La première solution est de ne pas harmoniser les mots-clés mais uniquement les noms d'ethnie et de lieu. On fait un profil commun avec les critères « titre », « auteur », « toponyme », « ethnonyme ». Comme au musée du quai Branly, on ne fait l'harmonisation que sur ces deux derniers champs en croisant les lieux et ethnonymes existants avec le vocabulaire Rameau. On peut ajouter le critère de recherche « sujet/mot-clé », mais en recherche plein-texte, sans harmonisation préalable. L'inconvénient de cette solution pour le champ « sujet » est un résultat approximatif. Une autre solution est d'accéder au contenu non par mot-clé mais par grand domaine comme il est possible dans la base Joconde du ministère de la culture.

[50] Direction des musées de France

3.1.2 Solution « portail Collections »

La deuxième solution est d'utiliser un moteur de recherche sémantique de type Lingway, Sinequa, Spirit. En amont, les algorithmes de moteur de recherche traitent les problèmes de langue à différents niveaux : syntaxique, morphologique, grammatical et sémantique (pour les synonymes notamment). En aval, le moteur travaille sur le corpus de résultats : tri par pertinence, reclassement thématique, clustering. Cela veut dire qu'on tolère d'avoir beaucoup de bruit dans les résultats si le système est capable de les placer en fin de liste et de faire remonter les réponses pertinentes.

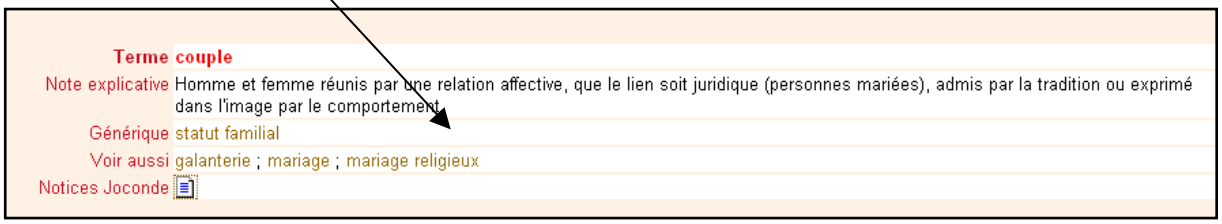
3.1.3 Solution « Le laboratoire d'idées d'Europeana »

La troisième solution est ce qui est développé par le web sémantique. Il s'agit de créer des liens entre différents thésaurus et lexiques documentaires : ces liens se font de manière automatique. Les équivalences de termes sont traitées en amont. L'inconvénient est encore une fois le bruit dans les résultats. Lorsque j'inscris le mot « mariage » dans l'encart de recherche et lance ma requête, une des réponses est *Scène d'intérieur avec couple âgé* d'Adriaen Van Ostade.

Europeana. Laboratoire d'idées. Scène d'intérieur avec couple âgé d'Adriaen Van Ostade

Dans la notice de ce tableau, les mots-clés sont : « couple, pot, scène, vieille…etc ». Le lien a été fait avec le mot-clé « couple » qui est rattaché au mot-clé « mariage » et « mariage religieux » dans le thésaurus Joconde.

Notice du thésaurus Joconde. Le terme couple renvoie à galanterie, mariage, mariage religieux.



Liens sémantiques du descripteur Mariage

Dans ce cas, le tableau n'a rien à voir avec le thème du mariage. La réponse n'est pas pertinente.

De même avec l'œuvre de Watteau *Le faux pas* dont on a parlé précédemment. Le terme **couple** apparaît dans les mots-clés de la notice. Un lien automatique est fait avec « mariage » car dans le thesaurus Joconde le terme **couple** renvoie à ce vocabulaire « galanterie, couple, mariage religieux ». Cependant, le tableau n'a pas de rapport avec le mariage.

Le faux-pas de Jean-Antoine Watteau

3.2 La documentation doit être accessible et utilisable en permanence pour l'édition en ligne

Il semble important de définir le public-cible et les objectifs du futur portail documentaire du MuCEM. Dans l'indexation du corpus Masculin-Féminin, nous nous sommes rapidement heurtés à ce problème. Les documents de ce corpus, s'ils étaient destinés à entrer dans le catalogue des collections du musée, devaient aussi être mise en ligne et donc s'adressaient à un public qui ne manie pas ce vocabulaire contrôlé.

Le thésaurus Garnier a une syntaxe contraignante et peu adaptée pour décrire une ressource orientée grand public. On a donc indexé les documents avec notre propre liste de mots-clés et simplifié les règles d'écriture. Avec Mondeca, une terminologie est en cours d'élaboration.

On pourrait créer deux champs « sujet ». L'un avec les mots-clés de la nouvelle terminologie grand public ; L'autre avec un vocabulaire contrôlé destiné aux professionnels et spécialistes. L'Ina (Institut national de l'audiovisuel) a choisi cette solution pour indexer ses documents.

Cette question d'accès à l'information via une recherche fédérée et harmonisation ou non des mots-clés, avec un thésaurus transversal ou non, peut être abordée sous un autre angle.

Le MuCEm est un musée de société qui pose des questions à ses contemporains, c'est-à-dire à tout type de public. Or celui-ci n'utilise pas nécessairement les outils documentaires tels que les portails documentaires et la recherche dans des bases de données. Il est possible de donner accès au contenu des collections sans passer par un portail de bibliothèque.

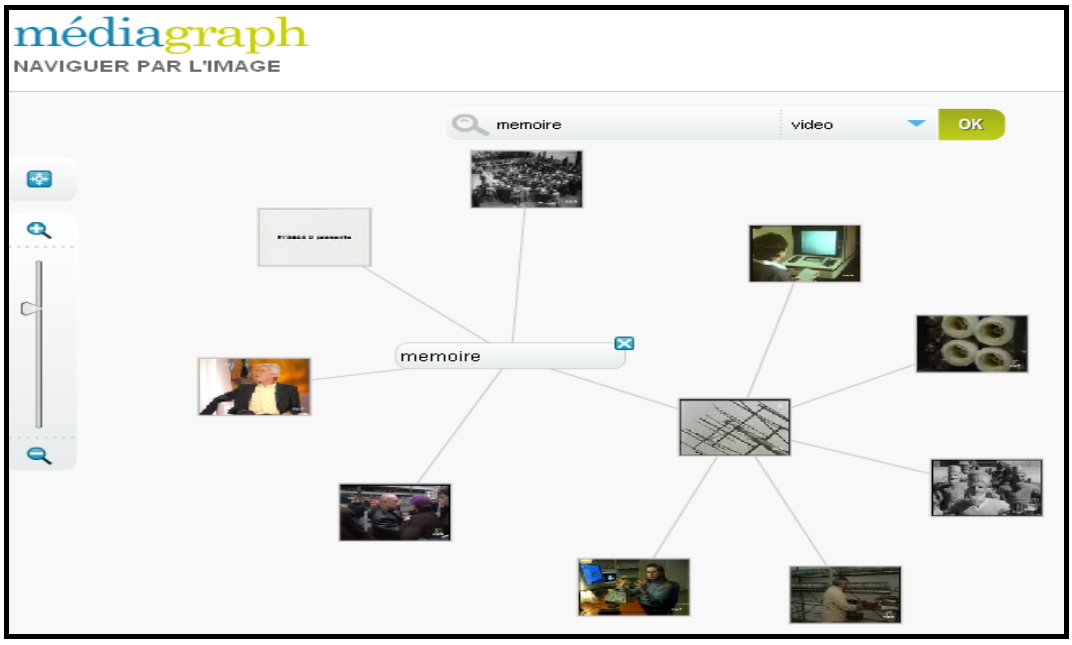
L'Ina s'adresse à la fois aux professionnels et au grand public. Pour exploiter ses ressources, elle met en ligne une sélection de vidéos à travers des frises interactives. Par exemple, cette fresque portant sur un siècle d'histoire permet de visionner des entretiens de penseurs, écrivains, artistes qui ont marqué leur temps. La fresque est classée par thème : culture, sciences et technique, relations internationales… etc. L'entrée thématique est souvent utilisée pour le public néophyte. Il facilite l'accès au contenu.

Entretien avec Louis-Ferdinand Céline. Accès à la vidéo et à sa notice descriptive très détaillée (notice, transcription, contexte historique)

Le structuralisme de Roland Barthes

L'Ina utilise aussi Mediagraph. En lançant une requête sur un sujet, l'usager accède à plusieurs documents qui renvoient à d'autres, via une interface cartographique. C'est une manière intuitive et facile pour l'usager de naviguer dans les contenus. Chaque vidéo est consultable et décrite par une notice détaillée.

Ces exemples montrent qu'il existe de nombreux moyens de valoriser les collections en les mettant en ligne tout en continuant le travail documentaire d'indexation de l'œuvre dans une base de données. Ce sont deux accès aux contenus différents et complémentaires.



INA. Mediagraph. Affichage des résultats de la requête « mémoire »

Chaque vidéo renvoie l'utilisateur à une autre, lui permettant d'explorer les collections de l'Ina sans passer par un formulaire de recherche simple ou avancé avec affichage de listes de résultats.

4 Le web sémantique

4.1 Définitions

Au cours de ce mémoire, différents termes ont été employés sans préciser leur définition. Or quelques notions ont besoin d'être clarifiées.

4.1.1 Le web sémantique et les moteurs de recherche sémantique

Le web sémantique et les moteurs de recherche sémantique sont deux technologies différentes. « Le terme web sémantique prête à confusion » comme l'explique Tim Berners-Lee dans un entretien [27, Berners]. Il préfère utiliser le terme de *web de données*.

La recherche sémantique ajoute à l'analyse statistique utilisé par les moteurs (poids des mots, nombre d'occurrences, calcul de page rank…) une autre couche basée sur l'analyse de la place et du sens des mots (analyse morpho-syntaxique et analyse sémantique). Ces technologies de recherche sémantique sont liées aux domaines du « text-mining » et du traitement automatique de la langue (TAL). Elles sont performantes sur des corpus homogènes en termes de vocabulaires et de structures des documents, mais restent limitées pour des corpus comme ceux proposés sur le Web. « Du coup, certains ont pensé qu'il s'agissait d'un Web qui permettrait par exemple d'effectuer des recherches sur Internet en posant des questions sous forme de phrases, en langage naturel. Or ce n'est pas son but. En fait, nous aurions dû l'appeler dès le départ Web de données. Mais il est trop tard pour changer de nom» [27, Berners].

Le principe du web de données consiste à relier toutes les données enfouies dans tous les ordinateurs de la planète. Actuellement, il faut utiliser des logiciels différents pour accéder aux données stockées dans les différents fichiers et bases de données de nos ordinateurs. Il s'agit de créer un lien automatique pour relier toutes ces données. « Ce lien offrira une interopérabilité inégalée », c'est-à-dire la possibilité de donner accès à ses données. « Si une entreprise met en oeuvre le Web sémantique, toute personne pourra accéder aux informations que cette entreprise a stockées sur ses produits et aussi aux informations stockées par d'autres entreprises sur les mêmes produits. Si quelqu'un cherche des photos sur un sujet et qu'il a besoin de récupérer le nom du photographe, les droits à payer, la définition de l'image etc., il accédera en une seule recherche aux photos et à ces informations, alors qu'avec le Web actuel il doit les chercher successivement dans plusieurs sites d'images. »

La finalité du web sémantique est d'améliorer les connaissances dans de nombreux domaines.

L'intelligence artificielle (IA) est la recherche de moyens susceptibles de doter les systèmes informatiques de capacités intellectuelles comparables à celles des êtres humains. Si le Web sémantique s'appuie sur certaines technologies mises au point dans le domaine de l'intelligence artificielle, à commencer par le principe des ontologies, ses ambitions sont plus limitées.

Le Web sémantique se caractérise par une orientation à destination des machines, à l'inverse des technologies Web traditionnelles (HTML, entre autres). En 2001, Tim Berners-Lee définissait le web sémantique comme « une extension du web actuel dans laquelle l'information reçoit une signification bien définie, améliorant les possibilités de travail collaboratif entre les ordinateurs et les machines »[51]. L'information sur le web actuel a pourtant une signification mais elle n'est accessible aujourd'hui qu'à des lecteurs humains. D'un point de vue d'informaticien, l'information est textuelle, peu structurée et donc inutilisable pour faire des traitements de calcul ou d'inférences.

The key enabler of the semantic web is the need of many communities to put machine-understandable data on the web which can be shared and processed by automated tools as well as by people. Machines should not just be able to display data, but rather be able to use for automation, *integration and reuse across various applications*[52] [38, Sure].

Lorsqu'on utilise un moteur de recherche, ce dernier n'est pas en mesure d'interpréter les informations contenues dans une page Web. Les technologies du Web sémantique vont permettre de mieux qualifier les informations mises à disposition sur le Web et ceci afin d'en permettre l'exploitation par les machines.

A l'inverse de la recherche sémantique qui s'appuie sur des algorithmes informatiques, les technologies du Web sémantique s'appuient sur une qualification explicite des données. Or, cette qualification est encore dans la très grande majorité des cas directement ou indirectement effectuée par des êtres humains.

---

[51] L'article fondateur du web sémantique, écrit par Tim Berners-Lee, James Hendler et Ora Lassila est *The semantic web : a new form of web content that is meaningful to computers will unleash a revolution of new possibilites.* Scientific American, mai 2001. http://www.sciam.com

[52] La clé de voûte dont dépend tout le succès du web sémantique est le besoin ressenti par tant de communautés d'afficher des données automatiquement déchiffrables (par des machines) sur le web, données qui peuvent être partagées et traitées aussi bien par des outils automatisés que des gens.

Pour résumé, le web sémantique ou web de données est l'idée est de construire un immense graphe (Giant Global Graph) qui relierait par le sens l'ensemble des données présentes sur le web. Le standard indispensable à la construction du web sémantique est l'URI (Uniform Resource Identifier = identifiant uniforme de ressource) qui permet d'identifier d'une façon certaine et sans équivoque une ressource. Le modèle de base du web sémantique est RDF (Resource Description Framework), «cadre» théorique et formel pouvant englober différents jeux de métadonnées (par exemple Dublin Core, XMP[53], SKOS[54], RDFS[55], OWL[56], FOAF[57]...) et dans des implémentations différentes (le plus souvent la syntaxe XML).

Exemples d'applications utilisant des nouvelles technologiques : Dbpedia et Wolfram Alpa

- DBpedia est un projet d'extraction de données de wikipédia pour en proposer une version web sémantique. Ce projet est mené par l'Université de Leipzig, l'Université libre de Berlin et l'entreprise OpenLink Software.

DBpedia est interconnecté avec GeoNames, MusicBrainz, CIA World Factbook, le projet Gutenberg et Eurostat, entre autres.

La base de données décrit 2 180 000 entités, incluant au moins 80 000 personnes, 293 000 lieux, 62 000 albums de musique et 36 000 films et contient 489 000 liens vers des images, 2 700 000 liens vers des pages extérieures, 2 101 000 liens vers des datasets externes et 207 000 catégories Wikipédia.Les informations étant stockées avec Resource Description Framework, on peut effectuer des requêtes sur la base de données via SPARQL. Le moteur d'extraction de données est réalisé avec PHP 5.

Leur but est d'extraire les informations de Wikipedia et de les rendre disponibles dans un format permettant des requêtes complexes sur des entrepôts de données constitués. La démarche est donc différente de Semantic mediawiki, puisqu'il s'agit ici de récupérer l'information déjà disponible dans Wikipedia. Les chercheurs de Dbpedia récupèrent les informations présentes dans les « infobox ». Les infobox rassemblent, sur le côté droit de l'article, un certain nombre d'informations de manière à peu près normalisée. Les chercheurs

[53] Extensible Metadata Platform ou XMP, format de métadonnées basé sur XML utilisé dans les applications PDF, de photographie et de graphisme.
[54] Simple Knowledge Organisation System (Système simple d'organisation des connaissances).
[55] RDF Schéma, langage extensible de représentation des connaissances.
[56] Web Ontology Language.
[57] FOAF (Friend Of A Friend), vocabulaire RDF permettant de représenter des informations sur les groupes ou les personnes.

ont extrait ces informations, les ont transformées en RDF et les ont intégrées à un entrepôt RDF.

« Les données de Wikipedia sont stockées dans une base de données relationnelles, mais, dans cette base, elles sont assez peu structurées, l'essentiel de la structuration étant assuré par le HTML. Un script aspire les pages de Wikipedia, les analyse automatiquement, dès qu'il trouve une « infobox », il transforme les informations de l'infobox en RDF/XML suivant une ontologie mis au point pour chaque type d'objet. Ces données en RDF/XML sont ensuite stockées dans un entrepôt spécialisé pour pouvoir les interroger en Sparql, le langage de requêtes de RDF »[58].

L'intérêt de cette technologie est de pouvoir faire appel très facilement aux données de Wikipedia depuis un autre site. Par exemple, dans un site, au passage de la souris sur un nom de personne, il sera possible de faire apparaître dans une info-bulle la date de naissance, la date de mort de cette personne. On pourra aussi interroger le dépôt en agrégeant les déclarations, c'est-à-dire les phrases simples sous la forme Sujet Prédicat Objet à la base de RDF. On peut aussi imaginer la construction de frise chronologique automatique à partir des données de Wikipedia grâce à Timeline, script mis au point par le Simile project. Ces exemples très simples ouvrent la voie à des mashups.

- Wolfram Alpha est un moteur de recherche « intelligent ». Il est capable de calculer, de comparer et de présenter synthétiquement des données disparates issues de bases de connaissances [39, Texier]. En ligne depuis mai 2009, Wolfram Alpha n'en est qu'à ses débuts comme le souligne son créateur Stephen Wolfram, scientifique d'origine britannique : « C'est la première étape d'un projet ambitieux : faire en sorte que toutes les connaissances soient calculables»[59]. Il ajoute que Wolfram Alpha n'est pas un moteur de recherche mais un « computational knowledge engine », littéralement « un moteur de savoir calculable ». Wolfram et son équipe ont récupéré des données appartenant à divers domaines (mathématiques, physique, chimie, ingénierie, géographie…) qu'ils ont structurées (mais pas selon les standards du W3C). Ces données sont calculées par des algorithmes dérivés du logiciel Mathematica, créé par Stephen Wolfram lui-même.

Une recherche sur « H20 » permet d'obtenir les principales caractériques de l'eau, y compris sa représentation moléculaire. Pour comparer l'hydrogène et le cobalt, il suffit de lancer la requête en anglais « hydrogen vs cobalt » et Wolfram Alpha fournit en quelques secondes

[58] Billet de blog, 11 février 2007. <http://www.lespetitescases.net/dbpedia-ou-la-puissance-du-rdf-au-profit-du-savoir>
[59] http://www.wolframalpha.com/about.html

une comparaison entre les deux éléments. Ses méthodes de calcul peuvent exploiter d'autres données. Pour comparer la population française à celle de l'Allemagne, l'utilisateur lance sa requête « population France vs population Germany » et les résultas s'affichent avec graphique, espérance de vie et autres caractéristiques.

4.1.2 Les ontologies

Cette notion apparaît dans les années 90 dans les recherches en modélisation des connaissances. Elle constitue la clé pour représenter explicitement et partager la signification véhiculée par les symboles informatiques. On peut la définir comme la représentation formelle et consensuelle des concepts propres à un domaine et des relations qui les relient. (Charlet).

Les ontologies sont les concepts de base du web sémantique qui cherche à s'appuyer sur des modélisations de ressources du web à partir de représentation conceptuelle. Le web sémantique a pour objectif de permettre à des programmes de faire des inférences sur ces représentations conceptuelles.

Aristote définissait l'ontologie comme cette partie de la Métaphysique qui spécule sur l'Etre en tant qu'être, indépendamment de ses déterminations particulières. Cette définition d'Aristote a-t-elle un lien avec celle étudiée dans le contexte de l'ingénierie des connaissances ? Cette question ne sera pas abordée ici [29, Charlet].

Une ontologie est un modèle des « choses qui existent ». L'objectif est de définir de façon formelle les concepts qui permettront de décrire ces « choses » de façon non ambiguë, et les règles contraignant ces descriptions. Ces concepts à décrire s'appliquent à un domaine de connaissances.

Une ontologie doit être compréhensible par les humains et utilisable par des machines pour des tâches diverses comme contrôler des interfaces, filtrer, classifier et agréger l'information, le cas échéant déduire de nouvelles informations [41,Vatant].

Dans leur article sur les ontologies, Charlet, Bachimont et Troncy donnent deux définitions de l'ontologie [29, Charlet]. La première décrit une ontologie comme « l'ensemble des objets reconnus comme existant dans le domaine ». Construire une ontologie, c'est décider de la manière d'être et d'exister des objets.

Dans cette définition, les objets ne sont pas pris dans un sens informatique mais comme objets du monde réel que le système modélise. Les ontologies sont développées dans un contexte informatique- que ce soit dans celui de l'ingénierie des connaissances, de l'intelligence artificielle ou du web sémantique. Dans ce contexte, l'ontologie devient alors un modèle des objets existants qui y fait référence à travers des concepts, les concepts du domaine.

La deuxième définition reprend les spécifications de T. Gruber et M. Uschold : « Une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts : entités, attribut, processus – leurs définitions et leurs interrelations. On appelle cela une conceptualisation. Une ontologie peut prendre différentes formes mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification ». [29, Charlet]

Tout programme informatique manipule, à travers des symboles, les objets du domaine modélisé. L'ensemble de ces objets correspond à un référentiel dans le domaine des systèmes d'information. Pour un domaine mettant en œuvre des connaissances complexes sur lesquelles on veut effectuer des traitements intelligents, le programme élaboré est un système qui manipule une base de connaissance. Celle-ci répertorie les concepts du domaine hiérarchiquement organisés dans une ontologie.

Les thésaurus et les ontologies procèdent de la même volonté de classer les choses. L'ontologie peut servir à l'indexation des documents au même titre qu'un thésaurus.
« J'ai une ontologie et pour chaque concept de l'ontologie, j'ai enregistré (dans le même fichier) le terme préférentiel pour nommer le concept et quelques synonymes : j'ai là ce qu'on appelle une ressource termino-ontologique (ou RTO). Et elle est totalement adéquate à une tâche d'indexation. Ce n'est pas le seul usage d'une ontologie, mais sous sa forme RTO, elle permet parfaitement ce travail ».[60]

4.2 Les fichiers SKOS : une solution pour harmoniser les mots-clés ?

4.2.1 Définitions

RDFS et OWL sont des langages d'ontologie qui permettent de décrire ce qu'est un document, les sous-types de document et les attributs spécifiques de chacun de ces types, y compris éventuellement le « sujet » du livre. Mais ils ne permettent pas d'exprimer

[60] Jean Charlet. Entretien.

l'organisation des vocabulaires décrivant les sujets. Ils ne peuvent pas représenter la hiérarchie de descripteurs, mots-clés ou catégories. SKOS va pouvoir le faire.

L'autopostage : Exemple de hiérarchie de concepts extrait du *Système descriptif des objets domestiques français* : Entretien>entretien-linge>lessivage>cuvier-à-lessive

Si on veut utiliser la hiérarchie de concept dans un but d'indexation, on ne peut pas utiliser OWL ou RDFS car cette hiérarchie de concepts n'est pas une hiérarchie de classes [41, Vatant].

Si j'indexe mon document avec le descripteur « lessivage », je dois pouvoir le retrouver avec des rubriques plus générales comme entretien ou entretien-linge. SKOS propose de répondre à cette question, celle de l'autopostage[61].

SKOS est un modèle de données permettant de gérer différents types de vocabulaires contrôlés tels que thésaurus, liste d'autorités, schémas classificatoires ou encore taxonomies. Il a pour vocation l'expression de concepts et de réseaux de liens conceptuels et sémantiques.

Normes et standard : Skos est passé, depuis le 18 août 2009, au statut officiel de « recommandation W3C ». Cette recommandation est pilotée par le Groupe de travail pour le déploiement du web sémantique (SWDWG, Semantic Web Deployment Working Group).

Une révision des normes internationales sur les thésaurus monolingues (ISO 2788:1986) et multilingues (ISO 5964:1986) a été lancée en juillet 2007 (document de novembre 2008).
Le 15 août 2009, le DIS (Draft international standard) sur les thésaurus a été publié.

L'identifiant et le titre de la (future) norme : ISO DIS 25964-1 - Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval (Information et documentation - Thésaurus et interopérabilité avec d'autres vocabulaires - Partie 1: Thésaurus pour la recherche documentaire)[62].

---

[61] « Procédé permettant d'effectuer automatiquement une indexation complémentaire d'un document ou d'une question par tous les descripteurs appartenant à la même branche de l'arborescence du thésaurus que le descripteur le plus spécifique utilisé lors de l'indexation. L'autopostage générique (vers un niveau supérieur) se fait lors de l'indexation ou lors de la recherche. L'autopostage spécifique (vers un niveau inférieur) s'effectue lors de la recherche ». Glossaire de l'ADBS

[62] Descripteurs : site dédié aux thesaurus et autres vocabulaires contrôlés pour l'accès à l'information<http://dossierdoc.typepad.com/descripteurs/>

La 2ème partie de cette norme est en cours. Elle portera sur l'interopérabilité entre thésaurus et avec d'autres vocabulaires : Thésaurus, Classification, vedettes-matière, taxonomie, ontologie, terminologie, anneau de synonymes.

4.2.2 Alignement et/ou correspondance des terminologies

Les alignements entre terminologies doivent pouvoir être publiés, distribués et réintégrés dans des applications, il est donc nécessaire de disposer de formats de sérialisation standardisés. Ce besoin a été pris en compte avec la norme SKOS (W3C), il existe donc aujourd'hui une possibilité de sérialisation normalisée en RDF/SKOS des données décrivant les alignements et mise en correspondance. La réutilisation et la distribution des alignements entre terminologies nécessitent l'utilisation des URI pour identifier les concepts mis en relation et permettre une réutilisation de l'alignement dans n'importe quel contexte par la suite [31, Delahousse ; 33, Macgregor].

A chaque concept peuvent être rattachés comme propriétés[63] :

- un terme préférentiel par langue
- des synonymes, avec spécification possible de la langue
- des définitions et des notes, avec spécification possible de la langue
- des concepts par des relations « générique –spécifique » ou par des relations associatives

Une fois un concept décrit par ces premiers attributs, il est explicité par ses relations à d'autres concepts à l'aide de la propriété *skos:semanticRelation* et ses propriétés dérivées. De la même manière qu'au niveau des attributs précisant les termes, ces propriétés sont particulièrement prévues pour représenter les relations sémantiques au coeur des thésaurus (hiérarchiques avec *skos:broader* et *skos:narrower*, et non-hiérarchiques avec *skos:related*).

Les propriétés de mise en correspondance proposées dans SKOS[64] permettent d'exprimer des correspondances (alignements exacts ou correspondances approximatives) entre concepts provenant de schémas différents[65].

---

[63] Rabault Hélène. SKOS : une ontologie des systèmes de représentation des connaissances. <http://semantiques.wordpress.com/tag/skos/>

[64] http://www.w3.org/TR/skos-reference/ - mapping

[65] W3C : the skos mapping properties

Les applications gérant des structures de concepts peuvent comparer les concepts et déclarer une identité de deux concepts par la propriété. Le schéma distingue une identité exacte *skos:exactMatch* (ex1:personne *skos:exactMatch* ex2:être humain) ou proche *skos:closeMatch*. Si la correspondance et la relation sont plus complexes, celles-ci sont exprimables par l'emploi d'un des trois attributs *skos:broadMatch*, *skos:narrowMatch*, *skos:relatedMatch*.

SKOS est un modèle de données commun pour le partage et la liaison des systèmes d'organisation des connaissances par l'intermédiaire du web [26, Bechhofer]. Ce standard et ses méthodes d'alignement automatique de vocabulaires pourraient être une solution pour faciliter la recherche d'information et la navigation dans des environnements hétérogènes.

Conclusion

Dans ce mémoire, nous avons tenté de comprendre comment la fonctionnalité de recherche fédérée des portails patrimoniaux met en valeur et donne accès aux contenus numériques ; plus précisément, grâce à quels outils et standards, les bases de données interrogeables lors d'une requête simultanée pouvaient devenir interopérables.

Il existe divers protocoles d'interrogation et le format Dublin Core permet de faire correspondre les champs des différentes applications. L'hétérogénéité des bases pose un véritable problème lorsqu' est abordée la correspondance des langages documentaires.

Les institutions culturelles ont choisi différentes solutions face à ces questions d'interopérabilité. Les solutions prenant en compte les technologies sémantiques semblent émerger. Le portail « Collections » utilise Sinequa, un moteur de recherche sémantique et « le laboratoire d'idées » d'Europeana a mis en place un alignement de ces terminologies grâce à une ontologie et au standard SKOS.

Sans confondre le web sémantique, tel que le définit son créateur Tim Berners-Lee, et les technologies de recherche sémantique, on pourrait affirmer que celles-ci et le standard SKOS répondent aux problèmes d'interopérabilité sémantique propres aux bibliothèques numériques.

Le MuCEM et la création de son portail documentaire étaient le point de départ de cette enquête. Grâce à la mise en place d'une bibliothèque numérique (Omeka) sur le serveur du MuCEM, il a été montré que la correspondance des champs des diverses bases de données au format Dublin Core n'était pas la principale difficulté. Celle-ci réside dans le fait que, sans thésaurus transversal, l'usager peut avoir des résultats de recherche non pertinents. Le MuCEM s'est associé à Mondeca pour créer une terminologie avec des outils linguistiques. Cette expérimentation pourrait couvrir par la suite l'ensemble des thématiques du musée et enrichir le vocabulaire actuel de façon semi-automatique. Pourquoi ne pas poursuivre ce projet et utiliser ITM, le logiciel de Mondeca, pour créer un portail sémantique grand public et gérer un langage documentaire transversal ?

Le site internet « Masculin-Féminin : histoire de couple et la construction du genre » permet de valoriser les recherches et les collections du musée. Le fait de les mettre en ligne oblige à faire évoluer la gestion documentaire de ces collections. La documentation du MuCEM doit être accessible et utilisable en permanence pour l'édition en ligne. Avec l'arrivée du numérique, les besoins et pratiques des publics ont évolué. Valoriser ces collections en le mettant en ligne nécessite d'assurer avec le plus grand soin le traitement « traditionnel » de l'information, à savoir l'alimentation des catalogues et inventaires des collections patrimoniales. Toutefois l'offre éditoriale, émanant de l'institution, et intégrant les

contributions des publics, fait partie des nouveaux modes de diffusion et d'échanges avec les visiteurs. La cohérence de la chaîne de traitement de l'information patrimoniale depuis la gestion documentaire jusqu'à la réalisation de produits de communication, en passant par l'édition électronique, passe par l'intégration et l'interopérabilité des différents outils utilisés dans une institution culturelle. Peu d'entre elles ont encore mis en place de tels dispositifs qui leur permettront d'accroître leur productivité intellectuelle et de remplir au mieux leurs missions de service public.

Bibliographie

La bibliographie a été arrêtée le 6 octobre 2009. Elle comprend 54 références.

La rédaction des références bibliographiques est conforme aux normes :

- Z44-005. décembre 1987. Documentation. Références bibliographiques : contenu, forme et structure et à la norme
- NF ISO 690-2 Février 1998 Information et documentation. Références bibliographiques Documents électroniques, documents complets et parties de documents

Bibliographie analytique et thématique

Cette bibliographie est classée par thème puis par ordre alphabétique d'auteurs. Les notices sont précédées d'un numéro entre **[ ]**. Ces numéros servent de référence dans le corps du texte. Les résumés qui accompagnent ces notices sont :

- Des résumés rédigés pour ce mémoire, signalés par la mention *(ag)*
- Des résumés d'auteurs, signalés par la mention *(r.a)*

Les notices sont organisées selon le plan de classement suivant :

1-Numérisation du patrimoine : Généralités

[1] LUPOVICI Catherine. Les usages des bibliothèques numériques : de Gallica à Europeana [en ligne]. **In** Les publics, 53e Congrès de l'ABF. Nantes, 8-11 juin 2007. [consulté le 10 juin 2009].
< http://www.abf.asso.fr/IMG/pdf/S5.1%20Lupovici%20mep.pdf>

Dans cet article, l'auteur présente le projet de bibliothèque numérique européenne, détaille les fonctionnalités de recherche, la classification utilisée, puis explique « le laboratoire d'idées », maquette mettant en œuvre un moteur de recherche sémantique et des ontologies. Pour finir, un point est fait sur les études d'usage qualitatives et quantitatives mises en place pour Europeana et Gallica. (ag)

[2] WESTEEL Isabelle. Le patrimoine passe au numérique. BBF [En ligne]. 2009, n° 1, p. 28-35. [Consulté le 10 juin 2009].
<http://bbf.enssib.fr/>

L'apport du numérique aux bibliothèques patrimoniales a bouleversé la valorisation, la communication et la diffusion des contenus et nécessite de modifier la gestion des collections pour prendre en compte les besoins de nouveaux publics. L'informatisation des catalogues, les plans de numérisation régionaux et nationaux, la recherche de visibilité dans les moteurs de recherche, le web collaboratif et sémantique, etc. sont autant de manifestations pratiques de ce bouleversement : le patrimoine passe au numérique. (r.a)

[3] Numérisation du patrimoine culturel / Dossier coordonné par Christophe Dessaux et Sonia Zilhardt. Culture et Recherche [En ligne]. Automne-hiver 2008-2009, n°118-119, 48 p. [consulté le 26 septembre 2009].
< http://www.culture.gouv.fr/culture/editions/r-cr.htm>

Ce dossier présente différentes réalisations de portails et services documentaires en Europe. Entre autres, le portail du ministère de la culture française Collections et la bibliothèque numérique Europeana. (ag)

2-Recherche fédérée, portails

[4] ARSENAULT Clément, PARE François-Xavier. Les portails de bibliothèque : nouvelles fonctionnalités, nouveaux défis. Argus [en ligne]. Automne 2005, vol. 34, n°2. p. 11–19. [consulté le 09 juin 2009].
https://papyrus.bib.umontreal.ca/jspui/bitstream/1866/136/1/Les%20portails%20de%20biblioth%c3%a8que%202.pdf

Les auteurs présentent des logiciels de portails de bibliothèque. La notion de portail de bibliothèque est définie, puis les principales fonctionnalités de ce type de produit (métarecherche, personnalisation, authentification). Un aperçu du marché des logiciels de portail est ensuite donné. Des questions reliées à la fiabilité des résultats et à la formation des utilisateurs sont également soulevées. (r.a)

[5] CERVONE Franck. Federated searching: today, tomorrow and the future (?). The Journal for the Serials Community [en ligne]. Mars 2007, vol. 20, n°1. p. 67 – 70. [consulté le 15 juin 2009].
http://uksg.metapress.com/app/home/content.asp?referrer=contribution&format=3&page=1&pagecount=

L'auteur donne des conseils pour mettre en place une fonctionnalité de recherche fédérée. Il ajoute que la sérendipité serait une valeur ajoutée à la recherche fédérée car elle permet la découverte de ressources jusque-là inconnues aux chercheurs. (ag)

[6] COLLIN Catherine, DISTEL Jeanne, MEURISSE Jack [et al.]. Présentation de "Collections", guichet unique d'accès en ligne aux données patrimoniales du ministère de la culture et de la communication. **In** COLLIN Catherine, DISTEL Jeanne, MEURISSE Jack [et al.]. Bases de données documentaires : état des lieux et perspectives, Journée d'étude, Ecole du Louvre, 22 mai 2007, Paris [En ligne]. [Consulté le 12 juin 2009].
<http://www.culture.gouv.fr/documentation/joconde/fr/partenaires/AIDEMUSEES/journee/journee-pres.htm>

L'auteur présente le portail « Collections » tant d'un point de vue technique que documentaire.(ag)

[7] GIBSON Ian, GODDARD Lisa, GORDON Shannon. One box to search them all : Implementing federated search at an academic library. Library Hi Tech. 2009, vol.27, n°1. p. 118-133.

Cet article aborde divers aspects de la recherche fédérée : les usages, les aspects techniques (les protocoles), l'affichage des résultats. L'auteur illustre ces propos par des exemples de moteur de recherche fédérée. (ag)

[8] MAISONNEUVE Marc. Le catalogue de la bibliothèque à l'heure du Web 2.0 : étude des opacs de nouvelle génération. Paris, ADBS, 2008. 306 p.

La première partie de l'ouvrage définit les concepts principaux, présente l'architecture et les composants de l'opac de nouvelle génération. L'auteur définit les termes de portails de bibliothèque et de recherche fédérée. La deuxième partie décrit dix solutions disponibles sur le marché français. (ag)

[9] SADEH Tamar. The challenge of metasearching. New Library World. 2004, vol. 105, n° 1198-1199. p. 104-112. ISSN 0307-4803.

Le service de recherche fédérée offre par l'intermédiaire du portail non seulement la découverte d'informations mais aussi la découverte de bases de données et de ressources documentaires.(ag)

[10] STILLER Henri. Le portail, outil fédérateur d'information et de connaissances. Documentaliste - Sciences de l'information [en ligne]. 2001, vol. 38, n°1, p. 39-42. [consulté le 26 septembre 2009].
http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2001-1-p-39.htm?WhatU=stiller%20henri&Auteur=&doc=N_DOCSI_383_0222.htm&ID_ARTICLE=DOCSI_381_0039

Cet article présente les principes, la typologie (portail internet, portail d'entreprise), les avantages et les difficultés de mise en œuvre des sites portails. (r.a)

[11] Pour ou contre la recherche fédérée ? **In** Encore un biblioblog. Billet du 8 novembre 2008. [consulté le 20 septembre 2009].
< http://bibliotheques.wordpress.com/2008/11/08/pour-ou-contre-la-recherche-federee/>

Après avoir distingué la recherche fédérée de la recherche unifiée, l'auteur analyse les usages des chercheurs en matière de recherche documentaire. Il explique que la recherche fédérée ne doit pas être un outil complexe à manipuler.(ag)

[12] Construire un portail, oui, mais, comment ? Journée d'études de l'ADBS-INTD. 21 octobre 2008. Mise en ligne le 24 octobre 2008. [consulté le 1er octobre 2009]. http://www.adbs.fr/interventions-construire-un-portail-oui-mais-comment--52550.htm

Etat de l'art du portail s'appuyant sur des exemples concrets de réalisations et introduit par deux exposés théoriques. (ag)

3-Normes, standards et langages documentaires

[13] FOULONNEAU Muriel. Collaborer pour de nouveaux services culturels en ligne : le protocole OAI, protocole de collecte de métadonnées de l'Initiative des Archives ouvertes. **In** Mission de la recherche et de la technologie. Relais, Culture, Europe. [En ligne], Janvier 2004, [Consulté le 12 janvier 2009].
http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/guide_oai.pdf

L'auteur explique le fonctionnement de l'OAI-PMH (collecte de métadonnées, moissonnage) et ses usages pour les services patrimoniaux.(ag)

[14] FOULONNEAU Muriel. Un Dublin Core Culture pour accéder à des ressources hétérogènes. **In** Mission de la recherche et de la technologie [en ligne]. Relais, Culture, Europe, avril 2003, [Consulté le 12 janvier 2009].
http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/dc_culture.pdf

Le Dublin Core fournit un ensemble de métadonnées de base pour rechercher des ressources hétérogènes d'origine différente et portant sur des sujets divers. Il comprend 15 éléments (Dublin Core simplifié). L'auteur rend compte des progrès de groupes de travail qui ont adapté ce standard au contexte culturel. (ag)

[15] Gauthier Mélanie. Le protocole OAI-PMH et les fonctionnalités de recherche : Etude de portails du domaine patrimonial. Mémoire de diplôme supérieur, Paris, INTD, 2007, 135 p.

Le protocole OAI-PMH a été rapidement adopté après son apparition par les institutions culturelles désireuses de créer des ponts entre collections et de se rendre plus visible sur le web par l'intermédiaire de portails permettant l'accès à leurs ressources disséminées. (r.a)

[16] LOIS Mai Chan, MARCIA Lei Zeng. Metadata Interoperability and standardization : a study of methodology, Part I. D-Lib Magazine [en ligne]. Juin 2006, vol. 12, n°. 6. [consulté le 15 juin 2009]. ISSN 1082-9873.
http://www.dlib.org/dlib/june06/chan/06chan.html

La croissance rapide des ressources internet et des collections numériques a été accompagnée par une prolifération de schémas de métadonnées. Chacun a été conçu en fonction des besoins particuliers des communautés, des futurs utilisateurs, des types de matériaux et des domaines propres…etc Les problèmes se posent lors de la construction de grandes bibliothèques numériques ou des référentiels de documents de métadonnées qui ont été préparés conformément à des schémas divers. L'auteur donne une définition du terme « interopérabilité ». (ag)

[17] NAWROCKI François. Le protocole OAI et ses usages en bibliothèque. **In** Ministère de la Culture et de la Communication [En ligne], Janvier 2005, [Consulté le 12 juin 2009]. http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm

L'auteur donne les principes fondamentaux du protocole OAI avec des exemples d'entrepôts et de moissonneurs. Des schémas illustrent les définitions. (ag)

[18] YOUSEFI Amin, YOUSEFI Shima. Metadata: a new word for an old concept. Library Philosophy and Practice, [en ligne]. 2007, Téhéran. [consulté le 15 juin 2009]. ISSN 1522-0222. http://www.webpages.uidaho.edu/~mbolin/yousefi.htm

La notion de métadonnée est utilisée dans le contexte des systèmes d'information moderne et les réseaux électroniques, remplaçant le terme de « catalogage ». Les métadonnées sont descriptives, administratives et structurelles. Plusieurs schémas de métadonnées sont énumérés, en particulier le Dublin Core considéré comme un standard. La difficulté consiste à utiliser un système de vocabulaire contrôlé global car même si les éléments des métadonnées communs sont utilisés, le contenu de ces éléments ne sont pas forcément compatibles. (ag)

[19] Dublin Core Metadata Initiative. [En ligne]. [Consulté le 12 juin 2009]
< http://dublincore.org/documents/usageguide/>

Guide d'utilisation pour les non-spécialistes : définitions, éléments, syntaxe, glossaire, bibliographie.

[20] EAD. Bulletin francophone de la direction des Archives de France sur l'EAD. [En ligne]. 2009, n° 35, janvier-mars 2009. [Consulté le 10 juin 2009].
http://www.archivesdefrance.culture.gouv.fr/static/2521

Ce bulletin informe des derniers évènements du monde archivistique. Il indique le guide de bonnes pratiques de l'EAD à la Bibliothèque du Congrès.(ag)

[21] The Open Archives Initiative Protocol for Metadata Harvesting. Open Archives. [En ligne], 2002, 7 juillet 2008. [Consulté le 28 septembre 2009].
http://www.openarchives.org/OAI/openarchivesprotocol.html

Version 2.0 du protocole, ce document donne les définitions et les concepts de l'OAI-PMH.

>> Langages contrôlés et système de classification

[22] GARNIER François. Thesaurus iconographique : système descriptif des représentations. [publ. par le] Ministère de la Culture, Direction du patrimoine, Direction des musées de France, Service informatique. Paris, le Léopold d'or, 1984, 239 p.

[23] CUISINIER Jean, DAVY de VIRVILLE, Michel. Système descriptif des objets domestiques français. Musée national des arts et traditions populaires. [s.l], Editions des musées nationaux, 1977, 291 p.

[24] The getty vocabularies : AAT, ULAN, TGN. [en ligne]. The J. Paul Getty Trust, 2009, [consulté le 12 juin 2009].
<http://www.getty.edu/research/conducting_research/vocabularies/>

[25] The icon class system. [en ligne]. The Netherlands Institute for Art History (Rijksbureau voor Kunsthistorische Documentatie, RKD), The Netherlands, 2006. [consulté le 12 juin 2009]
<http://www.iconclass.nl/

4-Web sémantique et outils de recherche sémantique

[26] BECHHOFER Sean, MILES Alistair. SKOS Simple Knowledge Organization System RDF Schema [en ligne]. 18 Août 2009. [Consulté le 20 septembre 2009].

http://www.w3.org/TR/skos-reference/

Ce document définit le Simple Knowledge Organization System (SKOS), recommandation du W3C et modèle de données commun pour le partage et la liaison des systèmes d'organisation des connaissances par l'intermédiaire du Web. Les notions de classe, propriétés et en particulier les relations sémantiques et de mise en correspondance des vocabulaires d'indexation sont explicitées. (ag)

[27] BERNERS-LEE Tim. Le web va changer de dimension. Propos recueillis par Marie-Laure Théodule [en ligne]. La Recherche, Novembre 2007, n°413 [Consulté le 12 septembre 2009].<http://www.larecherche.fr/content/recherche/article?id=6566>

Entretien avec le père fondateur du web qui donne sa définition du web sémantique : un web qui relie les données stockées dans différents fichiers et bases de données des ordinateurs. (ag)

[28] BURKE Mary. The semantic web and the digital library. Aslib Proceedings. 2009, vol. 61, n°3. p. 316-322. ISSN 0001-253X.

L'auteur propose une définition du web sémantique, explicite les relations entre web sémantique, web 2.0 et bibliothèque 2.0. Le monde des bibliothèques peut bénéficier de ces nouvelles technologies malgré un manque d'intérêt de la part des communautés des bibliothécaires pour le web sémantique.(ag)

[29] CHARLET Jean, BACHIMONT Bruno, TRONCY Raphaël. Ontologies pour le web sémantique. Information Interaction Intelligence. 2004, Hors Série. p.1-31.

Les ontologies sont un des concepts de base du Web sémantique. Les auteurs clarifient la notion d'ontologies, puis abordent les méthodologies de construction d'ontologies. Ils expliquent les apports du Web sémantique en termes de méthodologies, d'outils, d'éditeurs en essayant de préciser les problématiques particulières dans chaque domaine. (ag)

[30] CHARLET Jean, SZULMAN Sylvie, PIERRA Guy. DAFOE: A Multimodel and Multimethod Platform for Building Domain Ontologies. 2008. Journées francophones sur les ontologies. Lyon, France.<http://dafoe4app.fr>

Le concept d'ontologies, apparu dans les années 90, constitue la clé pour représenter et partager le sens porté par des symboles informatiques. Construire une ontologie est difficile.

Un moyen d'y arriver est d'utiliser des éléments préexistants : corpus de textes, taxonomies, normes ou autres ontologies, et de les prendre comme base pour définir l'ontologie du domaine. Il n'existe actuellement aucune procédure acceptée, ni un ensemble d'outils, permettant de concevoir de façon progressive, explicite et traçable une ontologie de domaine à partir d'un ensemble de ressources relevant de ce domaine. Cet article présente les propositions développées dans le cadre du projet ANR DaFOE 4app pour favoriser l'émergence de tels outils. (ag)

[31] DELAHOUSSE Jean. Sur l'alignement et la mise en correspondance de terminologies. In Mondeca. Leçons de chose [en ligne]. Paris. Billet du 29 juin 2009. [Consulté le 14 septembre 2009].
<http://mondeca.wordpress.com/2009/06/29/sur-l%e2%80%99alignement-et-la-mise-en-correspondance-de-terminologies/>

Cet article énumère les différents types d'alignement et de mises en correspondances répondant à divers types de besoins. Sont énumérés les moyens d'obtenir ou de créer un alignement ou de mettre en correspondance deux terminologies. L'auteur spécifie les types de relation décrivant ces alignements (relation d'équivalence, de hiérarchie, sémantique ou complexe), les standards de représentation et l'exploitation que l'on peut faire de ces alignements (alimentation des lexiques des moteurs de recherche, enrichissement de l'annotation des contenus, requête multilingues, mashup). Pour finir, il décrit les fonctions d'alignement et d'ontologie de la plate-forme ITM de Mondeca. (ag)

[32] GICQUEL Florence. Un glossaire peut cacher une ontologie...Documentaliste - Sciences de l'information, 2007, vol. 44, n°1, p. 64.

L'auteur présente la mise en place d'un glossaire multilingue pour un projet de gestion des connaissances entre la société de logiciel Mondeca et l'entreprise Lafargue, spécialisée dans le béton. La création de ce glossaire demandera la création d'une ontologie. (ag)

[33] MACGREGOR George. Introduction to a special issue on digital libraries and the semantic web : context, applications and research. Library Review. 2008, vol. 57. n°3, p. 173-177.

Pour beaucoup de bibliothèque numérique et institutions culturelles, le web sémantique offre une opportunité de mettre en valeur leurs ressources numériques et de résoudre le problème de l'interopérabilité, en utilisant des technologies et des standards communs d'une façon collaborative. Cet article explore cette question à travers des cas d'études et des

compte-rendus de recherches. Pour finir, le standard SKOS et les méthodes d'alignement automatique de vocabulaires sont mentionnés comme solution pour faciliter la recherche sémantique et la navigation dans des environnements hétérogènes. (ag)

[34] MENON Bruno. Les langages documentaires : un panorama, quelques remarques critiques et un essai de bilan. Documentaliste - Sciences de l'information, 2007, vol. 44, n°1, p. 18-28. ISSN 0012-4508.

Bruno Menon fait un panorama critique d'un siècle de développement et d'usages des langages documentaires ; puis examine les taxonomies et les ontologies. Pour conclure, il suggère un certain nombre de réflexions et de travaux qui permettront aux langages documentaires, dont le concept pourrait être étendu à celui de systèmes d'organisation des connaissances, de s'imposer demain dans l'économie numérique. (ag)

[35] MESGUISH Véronique. Le web sémantique : utopie ou réalité ? NETSOURCES, 2007, n°71, p. 1-5.

L'auteur définit le web sémantique, ses enjeux, le langage RDF, les ontologies et les pratiques d'indexation du web 2.0, et fait un panaroma des recherches en cours, entre autre les recherches de la société Mondeca et son logiciel ITM.(ag)

[36] MONDECA. Présentation [rapport interne]. 2009, 36 p.

Mondeca a pour mission de fournir logiciel et services pour aider les entreprises à organiser et valoriser leurs contenus. Cet éditeur, créé en 2000, utilise les technologies du web 3.0, ontologie et intelligence artificielle. Mondeca a créé ITM pour gérer les ontologies, les thésaurus, les taxonomies et les bases de connaissances afin d'indexer les documents.(ag)

[37] NORMIER Bernard. L'apport des technologies linguistiques au traitement et à la valorisation de l'information textuelle. Paris, ADBS éditions, 2007. 65 p. ISBN 978-2-84365-092-5

La question du traitement et de la valorisation de l'information textuelle est le thème de cet Ouvrage. L'auteur décrit l'utilisation de ces technologies à travers deux cas réels, l'un qui concerne l'accès à l'information sur les brevets et le second présente un outil de mise en correspondance de CV avec des offres d'emplois. (ag)

[38] RAIS Nadia. Principes et techniques d'indexation et de recherche de l'information textuelle. [Support de cours INTD]. Juin 2009, 33 p.

Après avoir rappelé des définitions de base de la documentation (information structurée, indexation et recherche), l'auteur donne les techniques du TAL et les typologies d'outils. Les différentes étapes de l'analyse linguistique (morphologique, syntaxique et sémantique) sont précisément décrites. (ag)

[39] SURE York, STUDER Rudi. Semantic Web technologies for digital librairies. Library Management. 2005, vol. 26, n°4/5, p. 190-196.

Cet article donne un aperçu général du web sémantique : historique, définition. Le terme « ontologie » est expliqué ainsi que les standards utilisés par le W3C. L'auteur précise en quoi les technologies du web sémantique peuvent aider les bibliothèques numériques.

[40] TEXIER Bruno. La recherche en langage naturel avance à pas de géant. Archimag, Septembre 2009, n°227, p.30-32, ISSN 0769-0975.

Cet article fait un état des lieux des moteurs sémantiques, de plus en plus utilisés par des logiciels profesionnels. Les internautes se plaignent du manque de pertinence des moteurs traditionnels. Ces moteurs apparaissent comme la panacée de la recherche en environnement numérique. L'auteur décrit Wolfram Alpha, une base de connaissances en sciences, capable de « comprendre » une requête.(ag)

[41] VATANT Bernard. Des métadonnées à la description des ressources : les langages du web sémantique. **In** Métadonnées : mutations et perspectives. Séminaire INRIA 29 septembre- 3 octobre 2008. Dijon, ADBS éditions, 2008. p.163-174

L'auteur explique les principes du langage RDF : triplet, graphe, URI, inférence. Il aborde ensuite les langages du web sémantique comme OWL et RDFS. Il donne une définition du terme « ontologie » et du standard SKOS qui s'impose comme le vocabulaire de référence pour la migration des vocabulaires d'indexation dans l'espace du web sémantique. (ag)

[42] Contre les idées reçus du web sémantique. **In** Les petits cases [en ligne]. Billet du 7 septembre 2009. [consulté le 12 septembre 2009].
< http://www.lespetitescases.net/contrer-les-idees-re%C3%A7ues-sur-le-web-semantique>

L'auteur donne une définition simple du web sémantique. (ag)

[43] Projet de recherche DAFOE 4App. [en ligne]. [consulté le 30 septembre 2009]. http://dafoe4app.fr/

Présentation du projet, de ses acteurs et des publications de chercheurs liés au projet DAFOE. (ag)

5-Musée

[44] BONTEMPI Julia. Quelle organisation documentaire dans un musée au défi des nouvelles technologies de l'information et de la communication : le cas du MuCEM. Mémoire de diplôme supérieur, Paris, INTD, 2007, 108 p.

Ce mémoire propose une réflexion sur les répercussions des nouvelles technologies sur la documentation muséale et son organisation. Après une première partie sur l'activité documentaire dans un musée, la seconde partie analyse la situation documentaire hétérogène du MuCEM et pour finir, donne des préconisations techniques (plate-forme logicielle, création de modèle de données, normalisation, procédure et modélisation). (r.a)

>>MuCEM : rapport interne

[45] BOELL Denis-Michel. Trésors du quotidien : l'Europe au musée des civilisations de l'Europe et de la Méditerranée. Avec la collaboration de Marie Robert et Dominique Vila. Paris, Réunion des musées nationaux, 2007. 162 p. ISBN 978-2-7118-5290-1

[46] COLARDELLE Michel, CALOGIROU Claire, BOELL Denis-Michel. Programme muséographique, 20 juin 2007. 514 p.

[47] COLARDELLE Michel, GERTREAU Florence. Du musée des traditions populaires au musée des civilisations, France. Nouvelles de l'ICOM, n°3, 2004, p. 6-7.

[48] DALBERA Jean-Pierre. Groupe informatique et multimédia : rapport d'activités 2008. MuCEM, Décembre 2008. 14 p.

[49] GIRARD Emilie. Le chantier des collections du MuCEM : présentation. 2007, 7 p.

>>Musée du Quai Branly : rapport interne

[50] GUILLOT Dominique. L'informatisation des collections au musée du quai Branly. Paris, Musée du quai Branly, janvier 2009. 34 p.

6-Masculin-Féminin, la construction du genre

[51 CHEVALLIER Denis, ZVEGUINZOFF Marina. Féminin-Masculin : projet pour l'exposition de référence. Thème 5 [rapport interne MuCEM]. Novembre 2007, 41 p.

[52] HERITIER Françoise. Masculin/Féminin I et II. Paris, Odile Jacob, 2002. 2 vol. (332-441 p.) Collection Bibliothèque. Index p.319. ISBN 978-2-7381-2041-0 et 978-2-7381-2040-3

[53] RAULT Françoise. L'identité masculine : permanences et mutations. Paris, la documentation française, 2003. 117 p. ISSN 0015-9743

[54] THERY Irène. La distinction de sexe : une nouvelle approche de l'égalité. Paris, Odile Jacob, 2007. 677 p. ISBN 978-2-7381-0984-2

Annexes

Annexe 1 : Le système d'information de la Bpi

Les entretiens avec les responsables de portail documentaire ont été nombreux. Certains ne figurent pas dans cette étude. Présentation du portail de la Bpi après un entretien avec Jérôme Villeminoz, Responsable de la Coordination bibliographique.

Un portail dans le site web :

Ouvert au public en avril 2006, le portail de la Bpi a été réalisé en partenariat avec la société Inéo media système. Il réunit sur une même interface l'ensemble des contenus produits ou acquis par la Bpi.

Le portail a d'abord été un site en tant que tel, indépendant du site web de la Bpi. En février 2009, il a été intégré visuellement au site web de la Bpi : il en est aujourd'hui la rubrique « Recherche documentaire ». Depuis cette intégration, le portail et le site sont toujours gérés sur des CMS différents. Le portail documentaire est géré avec Typo 3.

Accès au portail :

Il y a depuis le site web deux façons d'accéder au portail :

- un encart de recherche, présent dans le bandeau supérieur du site, permet de lancer une requête dans la catalogue de la bibliothèque ;
- un onglet « Recherche documentaire » donne accès au portail proprement dit.

Le système d'information actuel

Contenus :

Le portail donne accès, sur une même interface :

- au catalogue (OPAC Portfolio), intégrant tous les supports (livres, disques, films, ressources internet, archives sonores…)
- à 4 bases de données multimédias (BDM Médiaview) : bases de données, sites web sélectionnés, films documentaires, documents musicaux et sonores, documents d'autoformation.
- à la base Bpi-doc (GED), donnant accès à des articles de presse numérisés
- la bibliothèque numérique en cours de constitution, comprenant les archives du cinéma du réel et les archives sonores de la Bpi.

L'utilisateur accède aux ressources par l'interface propre de chacune des bases encapsulées dans l'interface unique.

Il existe différents modes de recherche pour accéder aux collections :

- accès par thème, via les « Pistes thématiques » : cette approche offre la possibilité de découvrir les collections multimédias et les documents de référence associés à un centre d'intérêt, ainsi que les produits documentaires réalisés par la Bpi (dossiers de presse, dépliants d'information, bibliographies d'actualité)

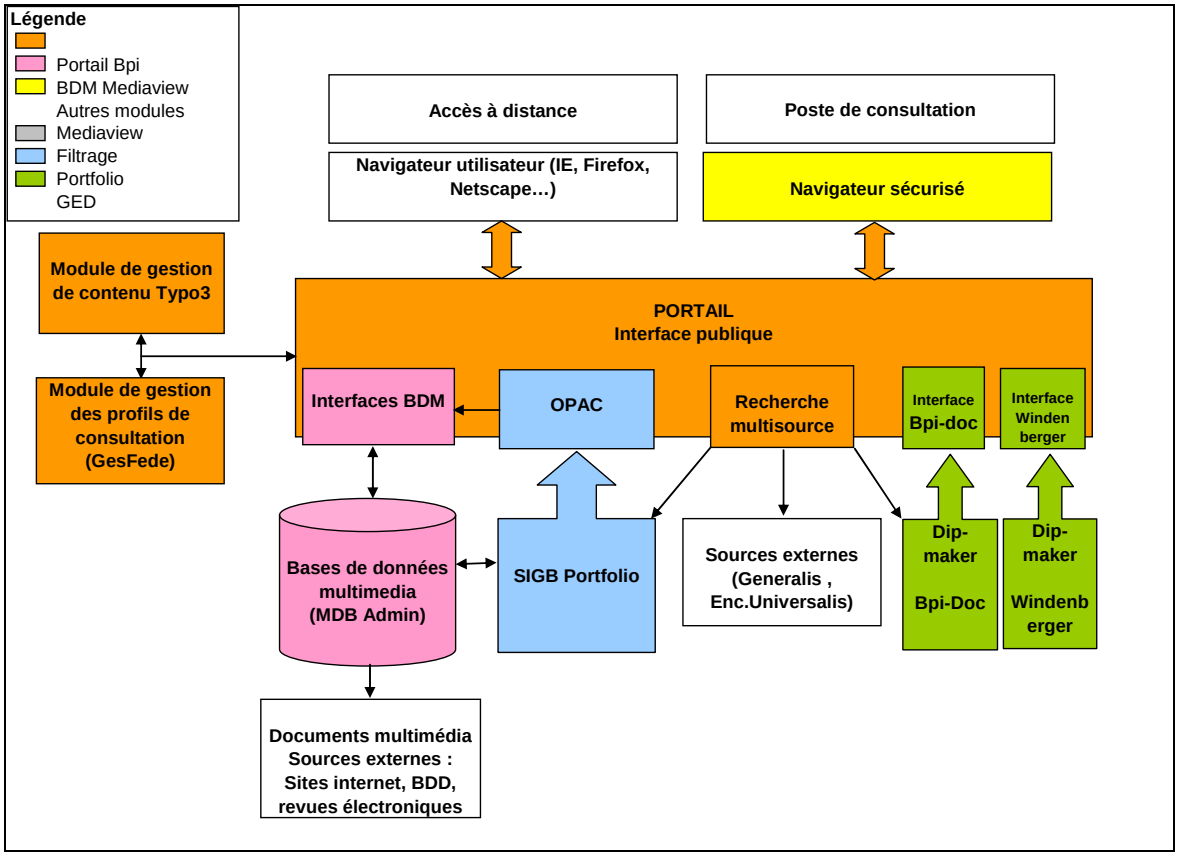
Accès à l'information par thème : Droit. Portail Bpi.

- accès par support, dans « Explorer par collection », ou en sélectionnant une base de données dans la rubrique « Ressources multimédias disponibles sur ce poste ».

- accès par le catalogue avec une recherche simple et avancée. On peut interroger les collections par les critères : auteur, titre, support, mot-clé.

- la recherche élargie : cet outil permet d'interroger simultanément le catalogue de la bibliothèque, la base Bpi-doc, les références d'articles de Généralis-Indexpresse et l'Encyclopædia Universalis. Ce n'est pas une recherche fédérée aboutie, dans le sens où les résultats ne sont pas fusionnés mais sont présentés base par base.

Les différents accès à l'information. Interface de recherche. Portail de la Bpi.



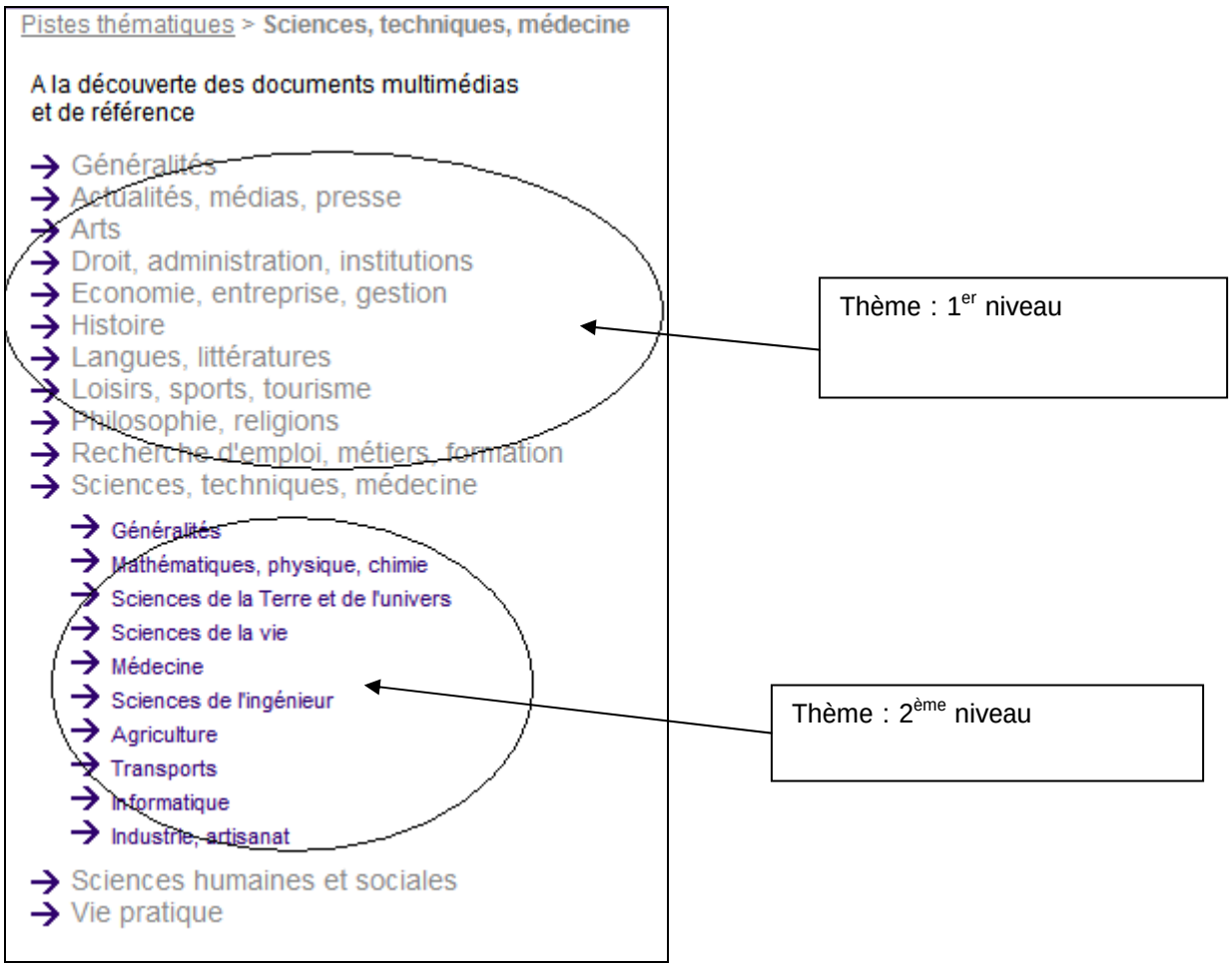
Le système d'information actuel de la Bpi

Harmonisation des mots-clés ? :

Concernant le catalogue de la bibliothèque, le langage d'indexation matière utilisé, quelque soit le support, est la liste d'autorité Rameau. Les documents sont organisés selon la Classification Décimale Universelle (CDU).

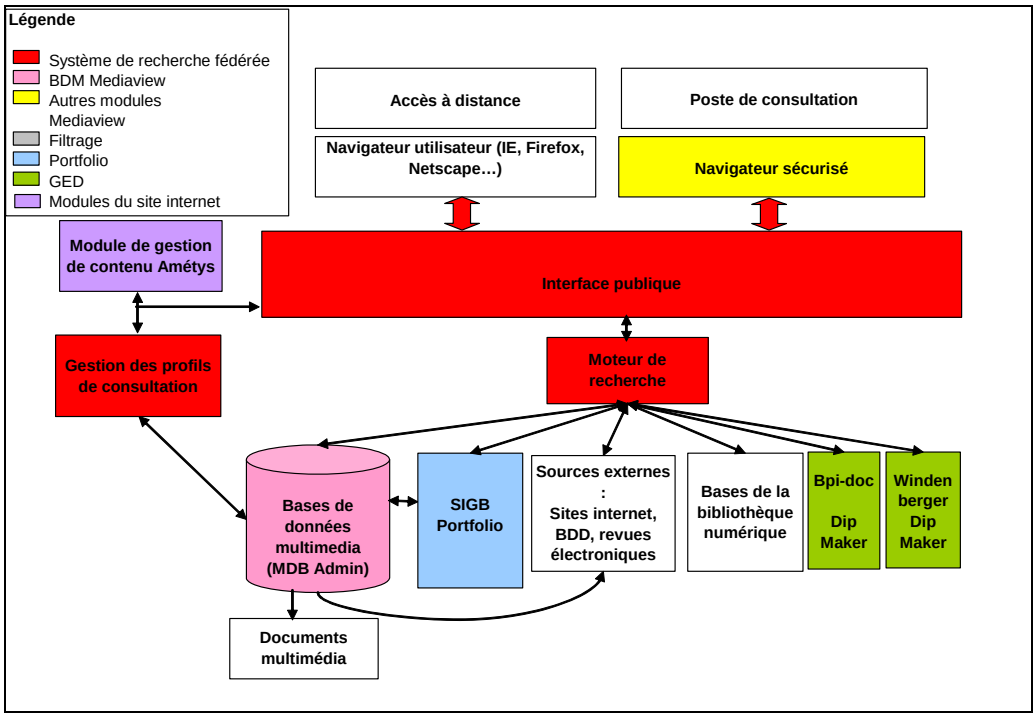
Pour l'accès par thème, un travail d'harmonisation a été effectué. En effet, les 4 bases multimédias (sites, films, son, autoformation), les BDD, les revues électroniques et les encyclopédies ont en commun un certain nombre de thèmes.

Cependant, chaque base a un choix de thèmes beaucoup plus large permettant une indexation plus fine. En effet, la classification thématique proposée s'organise sur 5 niveaux dont les deux premiers sont communs aux ressources indiquées plus haut et les trois autres propres à chaque base de données.



Classification thématique servant à l'indexation des documents

Le futur système d'information de la Bpi



Le futur système d'information de la Bpi

Le futur système d'information de la Bpi s'appuie sur la recherche fédérée, une approche des ressources transversale et non plus base par base, c'est-à-dire verticale.

Le moteur de recherche sera Library Find. C'est un moteur hybride, permettant soit de moissonner et de réindexer des bases de données (interrogation asynchrone), soit d'interroger directement leurs index (interrogation synchrone) .Ce moteur est en outre développé en *open source* par la bibliothèque d'état de l'Oregon.

Le moteur de recherche fédéré interrogera simultanément les bases de données produites par la Bpi (catalogue, Bpi-doc, bases de données multimédias etc) ainsi que la plupart des bases bibliographiques et sites internet gratuits ou payants proposés pour la recherche documentaire. L'objectif est de ne pas imposer à l'usager la compréhension de l'organisation technique de ressources documentaires, mais au contraire de lui proposer des logiques d'accès aux documents qui correspondent à ses besoins : au lieu d'avoir à interroger telle ou telle base, il pourrait chercher un document sur tel sujet ou un document de tel type.

Interopérabilité :

Dans LibraryFind, les métadonnées sont indexées au format Dublin Core (simple). Il reste donc à faire correspondre les différents champs de divers formats de chaque base en un

format Dublin Core. Concernant le contenu et la recherche par mot-clé, l'harmonisation des thèmes a déjà été faite sur une partie des ressources (collections multimédia, revues, documents de référence).

De plus, à moyen terme, la bibliothèque souhaiterait que l'index du moteur de recherche fédérée intègre les tables de renvoi correspondant aux formes retenues et formes rejetées des autorités du catalogue (fonctions "voir" et "voir aussi). Cela permettra d'étendre à d'autres bases de données le bénéfice du travail réalisé sur le catalogue. Exemple : dans Rameau, la forme retenue pour "municipalités" est"communes". Le projet serait que l'index du moteur de recherche fédérée intègre cette équivalence, pour que quand un usager tape"municipalités", le moteur étende la recherche dans toutes les sources en cherchant à la fois "municipalités" ou "communes".

Annexe 2 : Les sites web de la collection ethnographique du MuCEM

Les publications multimédias en ligne[66] sont :

1. **L'olivier, trésor de la Méditerranée** (en français et en italien) : les techniques de production, leurs évolutions, les symboles liés à l'olivier et à l'huile d'olive en Tunisie, France, Grèce, Italie et au Maroc. Un site réalisé en collaboration avec le Parco Nazionale del Cilento (Italie).

2. **Hip hop art de rue, art de scène** : l'histoire et les temps forts de l'expansion de ce mode d'expression propre aux nouvelles générations urbaines et qui a investi de nombreux domaines artistiques (musique, graphisme, danse, poésie). Le site prolonge l'exposition organisée à Marseille en 2005.

3. **Les voyages du verre** : les échanges de savoir-faire verriers dans le bassin euro-méditerranéen (République tchèque, Portugal, France, Autriche, Italie, Syrie, Pologne, Roumanie, pays arabes), les métiers du verre aujourd'hui (artisans verriers, ouvrières perlières, peintre sur verre).

4. **Café, cafés** : les objets, les modes et les lieux de la culture du café dans le monde, son économie et les techniques de torréfaction.

5. Cornemuses d'Europe et de Méditerranée : l'histoire de la cornemuse, de ses variantes dans les différentes cultures, de sa morphologie, de ses techniques de fabrication et du renouveau de sa pratique. Le site intègre un catalogue détaillé des 61 instruments conservés au MuCEM.

6. Les Petites Arménies d'Europe et de Méditerranée (en français, anglais et arménien) : l'histoire de l'Arménie, de son développement culturel et artistique et des diasporas arméniennes jusqu'au génocide au début du 20ème siècle.

[66] http://www.ethnologie.culture.fr

7. Les Trésors du quotidien, Europe et Méditerranée : La présentation de la totalité des objets de l'exposition de 2007 qui s'est tenue à Marseille avec la possibilité de téléchargement de l'audioguide.

8. Masculin-Féminin, histoire de couple et construction du genre : http://www.femininmasculin.culture.fr/

Glossaire

CDU	La classification décimale universelle (CDU) est un système de classification de bibliothèque développé par Paul Otlet et Henri La Fontaine, deux juristes belges fondateurs de l'Institut International de Bibliographie en 1895, à partir de la classification décimale de Dewey (CDD), et avec l'autorisation de Melvil Dewey.
Classement	Organisation de documents obtenus selon un ordre pré-défini (chronologique, alphabétique…) Elaboration de la classification : elle peut se faire en amont ou en aval de la recherche : organisation du fonds en catégories pré-définies ou regroupement des réponses en classes (clustering).
Classement automatique	Regroupement des documents dans des ensembles homogènes en fonction de leurs caractéristiques. 2 types de classification automatique : -catégorisation ou classement supervisé : classification de texte selon une classification préétablie : catégories définies au préalable (plan de classement, thesaurus) -clustering ou classification non supervisée : regroupement automatique d'un ensemble de documents présentant des similitudes en sous-groupes (clusters).
Clustering	Regroupement d'éléments dans des classes établies sur la base de similarités entre les documents.
Echange de données	L'échange de données consiste à faire communiquer de l'information entre différents systèmes plus ou moins hétérogènes. Par exemple, deux bases de données gérées par deux outils différents.
Index	(norme Afnor Z47-102) Un index est un ensemble ordonné de termes choisis et figurant dans un document avec une indication

	permettant de les localiser.
Indexation	Au sens documentaire : opération intellectuelle de description du contenu d'un document (mot-clés ou descripteurs) Indexation au sens informatique : opération automatique de création d'un index d'interrogation = fichier inversé
Indexation automatique	Représentation du contenu d'un document par sélection automatique de mots ou termes extraits de son texte, ou par attribution automatisée de termes extraits d'un langage documentaire. -Indexation en texte intégral : index constitué de tous les mots du document -Indexation assistée par ordinateur : repérage des concepts significatifs du document pour le caractériser avec des descripteurs validés ou non par un professionnel -Indexation par assignation : sélection des termes significatifs sur la base d'un langage documentaire
Information structurée	Information utilisable directement par un ordinateur pour effectuer un calcul. Ces calculs peuvent être variés : opérations arithmétiques (additions, multiplications…), comparaison (évaluation d'une requête booléenne par rapport à un document par exemple). Information stockée dans les bases de données. L'objet du traitement de l'information non structurée est de rendre des données (images, sons, textes) calculables. Ainsi ce n'est pas la nature de l'information mais l'usage d'une donnée qui marque la frontière entre structuré et non-structuré. (Livre blanc « Valorisation de l'information non structurée » Aproged _ Cigref. Oct. 2007)
Lemmatisation	Identification de la racine ou forme canonique pour pouvoir traiter les différentes variantes possibles.

Métadonnées	Ensemble d'éléments de données structurées afin de fournir des informations sur les ressources électroniques. Cette notion a remplacé celle de catalogage. Les métadonnées peuvent être : -Descriptives : description et identification des ressources : titre, source, date, volume -Administratives : gestion et conservation des documents : droit d'utilisation, droit d'auteur, cycle de vie, contrôle de qualité -Structurelles : pour la navigation et la présentation ; elles permettent d'établir des liens entre les documents, partie constituante du web sémantique.
Mot-clé	Mot choisi dans un titre ou le texte d'un document, sans référence à un lexique ou à un thésaurus, caractérisant son contenu et permettant la recherche de ce document. Le « mot-clé » diffère du « descripteur » : mot ou groupe de mots retenus dans un thésaurus ou un lexique de référence et choisi parmi un ensemble de termes équivalents pour représenter sans ambiguïté une notion contenue dans un document.
OPAC	Online public access catalog. Interface d'accès aux catalogues informatisés des bibliothèques
Sémantique	La sémantique s'intéresse au sens du langage pour en déduire des constructions logiques.
Syntagmes, identification des…	Groupe de mots qui se suivent avec un sens
Syntagmes, pondération des	Pondération des termes ou syntagmes : affectation d'un indice d'importance en fonction d'algorithmes de pondération.
TAL ou traitement linguistique de la langue	Reconnaissance de la structure des mots et des phrases : -Utilisation de ressources linguistiques (liste de mots vides, dictionnaires lexicographiques et terminologiques, lexiques, annuaires, fichiers référentiels d'identification d'entités nommées, corpus textuels mono pu multilingues, phraséologie, outils sémantiques : taxonomies, ontologies) -Analyse linguistique : morphologique (reconnaissance de la forme

	des mots), syntaxique (agencement de la grammaire), sémantique (signification de l'énoncé et sens lié au contexte)
Thésaurus	Selon la norme ANSI/NISO Z39-19, le thésaurus est un vocabulaire contrôlé utilisé pour la représentation de contenu dans des systèmes d'organisation de la connaissance. Il indique les termes utilisés pour décrire sans ambiguïté le contenu des documents au détriment d'autres termes, non utilisés. Le vocabulaire sur un thème précis est donc regroupé, limité, classé et régi par des relations de hiérarchie, d'équivalence ou de parenté entre les termes (descripteur générique, descripteur spécifique, descripteur associé). Les modalités d'emploi des descripteurs peuvent être brièvement expliqués dans des notes d'application (ou notes d'usage).
XML	Extensible Markup Language Les métadonnées normalisées basées sur XML sont entre autres : RDF Dublin Core : 15 éléments de catalogage minimal avec la possibilité de qualifier et de raffiner les éléments Biblio ML : transformation directe de la structure UNIMARC EAD : Encoded Archival Description : utilisé pour la description d'archives LOM.FR TEF : utilisé pour les thèses NewsXML : utilisé dans le milieu de la presse pour échanger des actualités ODF et Open XML pour le traitement de textes