

Thésaurus à la recherche

Sylvie Dalbin (sylvieatd@aol.com), Assistance & Techniques Documentaires

Document créé le 07/09/2007 – maj le 27/09/2007

Journée d'Etude ADBS - Paris, 20 septembre 2007

Optimiser l'accès à l'information, une opportunité pour les langages documentaires ?

Rappel des faits	1
1. Métamorphoses du contexte de la recherche documentaire	1
1.1. Métamorphose du monde documentaire	1
1.2. Démarche de recherche d'information : revue et corrigée	2
1.3. Des modalités d'accès à l'information qui se complètent	3
2 -Thésaurus à la recherche	4
En conclusion	9
Annexe – Rapport entre concepts, sujets et classes	10

Le passage à une Société dite de l'information ne se réduit pas à l'utilisation de technologies nouvelles, ni à leur simple intégration dans des systèmes documentaires préexistants. Dans les milieux professionnels, cette évolution impacte aussi bien l'organisation du ou dans le travail, que les besoins et les pratiques des collaborateurs dans l'utilisation des systèmes d'information (pour trouver), et in fine de l'information elle-même.

Dès lors comment aborder les thésaurus de descripteurs¹ ?

Après une période de quasi-monopole dans les bases de données professionnelles, puis une deuxième période au sein des bases documentaires des centres documentaires, les thésaurus de descripteurs ont été, si ce n'est jetés aux orties, à tout le moins oubliés dans les années 90 en France pour renaître de leur cendre depuis quelques années.

Le contexte et les besoins ayant évolué, comment aborder aujourd'hui ces thésaurus, et plus particulièrement le thésaurus à la recherche ?

Rappel des faits

Le thésaurus a été largement utilisé à la recherche d'information en miroir du thésaurus à l'indexation : les documents indexés avec un thésaurus étaient utilisés avec le même thésaurus à la recherche. Pourtant dès le début des années 1980, les spécialistes des thésaurus anglo-saxons explorent ce que pourrait être le « thésaurus Utilisateur » (user thesaurus²) ou le « thésaurus à la recherche »³, en précisant les caractéristiques de ce type d'outils : un Superthésaurus² ou encore des listes de synonymes, des réserves de termes, des thésaurus traditionnels élargis, des thésaurus fusionnés avec des traits particuliers, des thésaurus indépendants du thésaurus à l'indexation² (pp.150-151). Ces propositions reposaient à l'époque sur le développement des pratiques de recherche documentaire en ligne et sur la diversité des profils d'utilisateurs. Sur ce dernier point, la situation avec les internautes, s'est largement intensifiée. L'autre élément-clé de ces propositions portait sur la nécessité d'une emprise plus grande de l'informatique. C'est sur ce deuxième aspect – la prise en charge du thésaurus par la machine et non l'utilisateur, que les changements sont les plus radicaux.

1. Métamorphoses du contexte de la recherche documentaire

1.1. Métamorphose du monde documentaire

Le paysage documentaire évolue fortement repositionnant le rôle des thésaurus et plus globalement celui des vocabulaires contrôlés. Parmi les critères ayant des conséquences directes

¹ Nous utilisons la dénomination « thésaurus de descripteurs » ou thésaurus documentaire, pour les thésaurus exploités dans des bases documentaires pour les dissocier des thésaurus de langue (Roget's thésaurus créé en 1870) très utilisés dans le monde anglo-saxon, d'où d'ailleurs est issu le thésaurus de descripteurs.

² How to Use Controlled Vocabularies More Effectively in Online Searching, Marcia Bates, 1989, <http://www.gseis.ucla.edu/faculty/bates/>

³ Construire un thésaurus (1987), Jean Aitchinson et Alan Gilchris, ADBS Edition, 1992, pp.150-151., Lancaster (chap.17) ; Searching Vocabularies, Piternick, Anne B., *Online Review* 8, no.5, pp. 441-9, 1984.

sur les thésaurus, on peut citer :

- Une augmentation des pratiques d'acquisition de ressources et de métadonnées associées à ces ressources, avec comme corollaire une baisse notable des activités d'indexation au profit d'autres activités en amont (sources) ou en aval (offre). [*Tout le monde n'indexe plus ou n'indexe plus autant : la place du thésaurus dans les pratiques des professionnelles est remise en question*].
- Des pratiques d'extension de l'offre documentaire à d'autres ressources accessibles en interne ou à distance tels les réservoirs types OAI ou z3950/SRU. [*Les ressources sont traitées selon des pratiques variables et avec des langages différents.*]
- Le déploiement des documents numériques permettant un traitement des contenus des ressources elles-mêmes en complément des métadonnées. [*Une richesse, pouvant être aussi source de bruit*]
- Le développement de filières d'indexation ou de classification automatiques de documents pour lesquels une indexation humaine ne semble pas pertinente en termes de coût et d'intérêt.
- Des espaces documentaires sur les intranets/extranets, volumineux, multiformes, avec une tendance à l'enrichissement en métadonnées directement à la source par les producteurs ou les auteurs, et des possibilités d'interaction en aval avec des lecteurs (annotations).
- Des documents multiformats (texte, image, son...) et de plus en plus souvent plurilingues, pour lesquels l'indexation se pose en termes assez différents.

Des fonds mobiles et des ressources décrites selon des modalités variables sont les caractéristiques principales qui remettent non pas en cause mais en question la place du thésaurus dans les systèmes documentaires.

1.2. Démarche de recherche d'information : revue et corrigée

1.2.1. Multiplicité des profils d'utilisateurs et de leurs pratiques

A côté des chercheurs scientifiques, qui restent bien trop souvent encore le modèle de référence dans la conception des systèmes documentaires, la Société de l'information draine de très nombreux autres profils d'utilisateurs.

Dans l'environnement professionnel, ces utilisateurs peuvent avoir à conduire des recherches d'information soutenue et régulière (veilleurs, chargés d'étude), mais leurs besoins peuvent être plus ponctuels, plus diffus. Pour ceux-là, l'ergonomie des interfaces devient cruciale et il faut bien le dire l'appréhension d'un thésaurus de descripteur devient problématique.

Les situations de travail ont aujourd'hui des caractéristiques que l'on peut qualifier de nouvelles sur le plan de l'accès à l'information :

- L'activité professionnelle de ces utilisateurs n'est pas centrée sur la recherche d'information mais sur une production autre ; la finalité de cette recherche d'information peut-être très variée : constituer un dossier ou rédiger une note, mais aussi répondre à une sollicitation ponctuelle, compléter un paragraphe dans un document collectif, alimenter un autre système d'information...
- La recherche porte sur des informations de nature très variable : un fait, des données statistiques ponctuelles, une explication, une norme, un formulaire, des images, la décision prise 3 ans auparavant sur un point particulier, un manuel d'utilisation d'un appareil, un article sur un sujet non connu, un paragraphe ou un schéma issu d'un article précis, ...
- Les fonctionnalités attendues des systèmes documentaires sont tout aussi variés : extraire une donnée statistique ponctuelle ou un fait d'un document pour l'intégrer dans une note, fouiller 1500 documents sur un thème précis pour alimenter sa réflexion et proposer des pistes de travail, ou pour comprendre les données d'un problème (outils de fouille de données), suivre l'écho dans la presse de tel produit (fil RSS), élaborer les prévisions budgétaires du service (des documents de référence) en contrôlant les procédures dans la dernière version du manuel comptables,...
- Une masse de sources à exploiter : une variété de logique et de modalités de consultation

Les moments dévolus à une tâche de recherche d'information restent très nombreux et surtout variés dans leur nature, mais elles sont porteuses de contraintes fortes et ne se résument pas au lancement d'une requête sur un portail ou une base documentaires. On pourrait résumer le besoin comme ceci : être informé en faisant le moins d'effort (en temps mais surtout en effort cognitif), et

si possible sans même avoir à effectuer une recherche !

1.2.2. Les utilisateurs face au système

Le rapport que ces utilisateurs entretiennent avec leur propre questionnement (clair, flou, ambigu, précis...) et avec les ressources interrogées (connues, inconnues) constitue une variable qu'il convient de bien étudier lors de la conception des interfaces d'accès à l'information. On peut très schématiquement identifier 4 situations extrêmes :

- Le modèle initial lié au « catalogue de bibliothèque » : besoin documentaire bien précisé par rapport au problème posé et correctement exprimé par rapport aux fonds/documents. Il s'agit plus de retrouver quelque chose que l'on connaît, dont on connaît l'existence ou a minima dont on connaît des éléments de sa description.
- Proche de ce modèle : besoin mal précisé mais fonds connu. L'utilisateur est à même d'exploiter des index ou une arborescence thématique qui font sens pour lui ; partant de ce premier point d'ancrage, il peut étendre sa recherche par rebond ou au contraire l'affiner.
- Plus éloigné de ce modèle : besoin bien précisé mais fonds mal connus ou inconnus d'où des difficultés à exprimer une requête au système. Une prise en charge de la question par le système est ici nécessaire. Une première requête ouverte doit aboutir à une présentation des résultats de type tableau de bord à même de guider l'utilisateur pour rebondir vers un ailleurs ou au contraire pour permettre de raffiner la question
- Encore plus éloigné : besoin imprécis, idée floue, pas ou faible connaissance sur la nature d'information susceptible de réponse à son besoin, ou bien vision partielle. Ici, un schéma de représentation du système interrogé permettra à l'utilisateur d'initier une requête pour aboutir au cas précédent.

1.3. Des modalités d'accès à l'information qui se complètent

Si les situations des utilisateurs face aux systèmes sont variées, la panoplie des modalités d'accès à l'information est somme toute assez réduite. C'est plus leur mise en œuvre et leur articulation au sein d'interfaces qui font la différence.

Quelles sont ces modalités générales d'accès à l'information ?

1.3.1. Sélectionner un item au sein d'une liste de thèmes ou de sujets proposés

L'utilisateur trouve au sein de ces listes, un élément représentant totalement ou partiellement son sujet : type FAQ (foire aux questions), répertoire de DSI (diffusion sélective de l'information) prêtes à l'emploi, index ou encore liste de thèmes⁴. Ces modalités sont tout à fait d'actualité même si l'on peut noter, en concordance avec le peu de pratique des index de fin de livre ou de site Web en France, un usage moins fréquent de ces modalités en France par rapport au monde anglo-saxon.

1.3.2. Naviguer dans une représentation des connaissances

La construction de cette représentation est orientée par les connaissances et besoins des usagers, et non par la représentation des connaissances académiques type de celles portées par les classifications Dewey ou CDU. Il s'agit ici de plans de classement spécialisés ou plutôt contextualisés par rapport aux activités. Le développement de certaines ontologies métiers, taxonomies⁵ ou plans de classement utilisés pour les systèmes de records management entrent dans cette catégorie. Dans certains cas, ces taxonomies peuvent ressembler aux thésaurus tels que nous les connaissons depuis 40 ans ; en fait ces schémas de navigation correspondent aux *catégories*, liste des champs sémantiques ou microthésaurus des thésaurus, mais ne proposent

⁴ De 189 sujets (<http://www.library.okstate.edu/Govdocs/browsetopics/>) pour le portail de sites gouvernementaux américains, à 21 grands domaines pour Service-Public.fr, ou 35 catégories sur e-Bay. Autres exemples d'index de sites Web : Université : <http://www.carlisle.mec.edu/A-Z-index.html>; Réserve fédérale USA: <http://www.federalreserve.gov/azindex.htm> ; Australie: http://www.aph.gov.au/find/find_index.htm ; Santé : <http://www.health.state.ny.us/healthaz/> (en anglais et espagnol).

⁵ Un exemple : Des taxonomies dans le domaine de l'éducation en sciences <http://pratic.ens-lyon.fr/projets/meta-LOM-CDM/thesaurus/thesaurus>.

pas les centaines de concepts eux-mêmes qui composent le graphe du thésaurus (voir annexe).

1.3.3. Formuler soi-même une question à partir de mots clés

La formulation de la requête doit être simple et le système doit aider l'utilisateur en aval, au filtrage du lot résultat ou à la navigation au sein de ces résultats.

Il convient de noter que ces différents modèles de représentation des sujets des documents ou des questions ne s'excluent pas, et qu'il s'agit dès lors de trouver le meilleur compromis entre une représentation des contenus avec des descripteurs ou des listes de sujets, une organisation de ces contenus avec une taxonomie ou une classification, et l'assistance d'un moteur de recherche intelligent.

Un exemple : Le site de l'INRS (voir l'intervention faite au cours de cette même journée d'étude), et au sein de ce site, le Guide électronique d'accès aux Tableaux des maladies professionnelles proposé par l'INRS (Rubrique Base de données de l'INRS : <http://www.inrs.fr/>) Régime (agricole, général), Pathologie (plan de classement), Pathologie (liste alphabétique), Agent ou nuisance en cause, Travaux effectués, Numéro de tableau et Texte intégral.

2 -Thésaurus à la recherche

Entre les pratiques des utilisateurs, les techniques informatiques et les espaces documentaires en mutation, comment situer le thésaurus à la recherche ? Nous proposons de suivre les étapes qui jalonnent une activité de recherche d'information, pour pointer à quel moment un tel outil pourrait être utile et utilisé.

#1. Où dois-je aller ?

De son bureau électronique (et physique), comment l'utilisateur procède-t-il ! On peut « obliger » l'utilisateur à aller sur un système dédié : soit l'intranet, soit un espace spécialisé (veille, bibliothèque, revues électroniques, coin presse...), soit au sein de ces espaces un fonds précis pour lequel une interface particulière aura été développée. Cette première étape conditionne bien sûr la suite. Nous ne l'aborderons pas ici, mais elle reste un point noir du processus pour l'utilisateur, encore trop souvent oublié⁶.

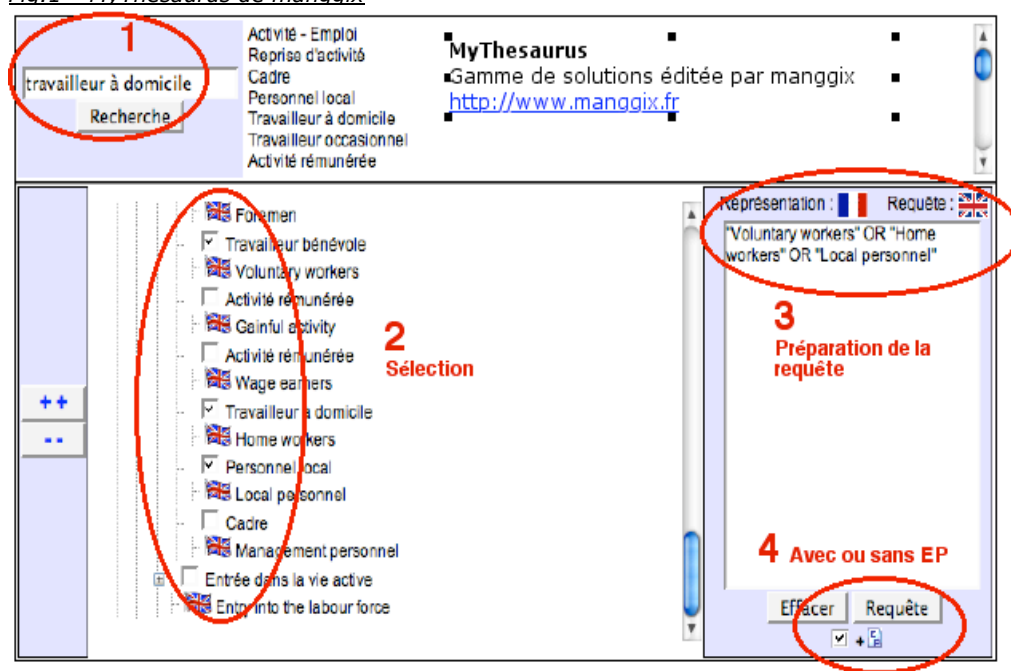
#2. Comment formuler ma requête ?

L'interface dédiée à la formulation précise d'une requête reste toujours d'actualité, même si elle n'occupe la première place. Il s'agit ici de proposer un formulaire permettant la sélection de termes au sein des index ou d'un schéma de représentation des connaissances type thésaurus. Ce type d'interface peut être proposée pour interroger des ... moteurs de recherche.

- Un thésaurus peut être exploité comme un dictionnaire pour sélectionner les termes de sa requête. C'est l'usage à l'origine des thésaurus de mots. Ceci est particulièrement pertinent dans le cas où vous voulez interroger des ressources dans une langue étrangère pour vous ou un secteur dont vous ne connaissez pas la terminologie. L'écran suivant (Fig.1) montre MyThésaurus de manggix, un outil permettant d'utiliser un thésaurus comme réservoir terminologique pour interroger des moteurs de recherche.

⁶ L'exploitation d'un moteur de recherche dans une première étape, peut conduire l'utilisateur directement à une page précise d'un espace documentaire.

Fig.1 – MyThésaurus de manggix



La requête élaborée en 3 à partir d'un thésaurus (avec ou sans les équivalents en 4) est alors lancée sur un ou plusieurs moteurs au choix de l'administrateur.

- Des interfaces spécifiques peuvent intégrer des métadonnées que l'utilisateur sélectionne avant de lancer sa recherche et qui servent de filtres à une recherche libre. L'architecture du dispositif s'appuie alors sur des métadonnées bien précises : les plus importantes pour les utilisateurs cibles en rapport avec leurs activités ou alors une sélection parmi les métadonnées proposées par des schémas spécialisés comme le LOM-Fr pour les ressources à caractère pédagogique (public, type de ressources pédagogiques...). Le principe est d'isoler des métadonnées en fonction de l'audience : catégories de structure d'accueil pour le secteur touristique, dans le monde du sport événements - noms de sportifs - sports ; dans le monde de l'assurance, produits d'assurance,... ; nom de personnage en littérature, période et zone géographique en histoire, L'architecture des intranets repose souvent sur cette approche, dite par abus de langage, à facettes. Associées à ces métadonnées, on trouve des listes contrôlées dans lesquelles l'utilisateur peut piocher. Ces listes correspondent bien souvent aux listes dites annexes intégrées aux thésaurus et que l'on peut en grande partie associer aux entités nommées : listes de noms de personnes, de sociétés, de lieux/zones géographiques mais aussi listes de noms de produits, de marques, de sports... Un travail de mise en cohérence entre ces ressources terminologiques serait intéressant à réaliser.

Dans cette formule où l'on isole des listes de termes, le thésaurus vu comme un tout interrogeable, est en quelque sorte déstructuré.

Nous pouvons noter que les autres catégories de termes d'un thésaurus, les noms communs, ne nous semblent pas appropriés à une exploitation de type navigation selon les logiques classificatoires : par principe, un seul terme sélectionné ne correspond qu'à une partie seulement d'un sujet, et le système renvoie alors de très nombreuses réponses qui peuvent être assez éloignées du sujet. Le schéma proposé en annexe expose les relations entre les sujets des questions, et les termes d'un thésaurus ou les classes d'un plan de classement.

Mais le souhait des utilisateurs de formuler leur requête au plus simple au lieu d'utiliser des formulaires ou des équations de recherche complexe impose pour assurer le même niveau de qualité aux résultats fournis, de travailler soit sur le traitement de la requête (3), soit sur le traitement du lot résultat (4).

#3. Comment la machine traite-t-elle la requête (en back office)

Les techniques informatiques de ces 20 dernières années permettent d'envisager l'exploitation non seulement des index des bases, mais surtout du thésaurus lui-même de façon intelligente.

3.1. Une liste de synonymes pour enrichir les termes de la requête

Une des premières exploitations de thésaurus, exploitation citée d'ailleurs par Lancaster dès 1984, est celle des listes de synonymes. Il s'agit d'enrichir la requête par des termes synonymes ou par des équivalents. Moins contraignante que l'exploitation du graphe complet du thésaurus, cette fonction permet de reformuler et d'étendre les requêtes des utilisateurs. Au milieu des années 1990, la plupart des outils de TALN (traitement automatique de la langue) se sont concentrés sur cette fonctionnalité de reformulation de la question : à un coût moindre, les fonds n'étant pas nécessairement indexés avec le même système, la qualité des résultats est ainsi considérablement améliorée. Ce type de montage est assez simple à imaginer, mais malheureusement les modalités de recherche dite simplifiées - c'est-à-dire à la « Google » - proposées par les portails documentaires bien souvent n'intègrent pas ce traitement minimal de la requête ce qui nuit grandement à la qualité de la recherche⁷.

3.2. Un thésaurus pour relancer une recherche

Dans le cas présenté ci-après, la première requête reste libre selon les fonctionnalités du moteur interrogé mais il est possible d'exploiter un thésaurus ou plusieurs thésaurus pour reformuler et étendre une recherche.

Exploredge propose une interface dédiée où peuvent être embarqués plusieurs vocabulaires. Sur cet écran (Fig.2), l'interrogation d'un moteur comme Exalead montre la pertinence de l'utilisation de thésaurus proposant des termes spécifiques ou génériques n'apparaissant pas dans les résultats du moteur. Les termes du thésaurus viennent compléter les propositions thématiques faites par le moteur et pour ces dernières, extraites automatiquement du lot résultat de recherche.

Fig.2 - Exploredge

The screenshot shows the Exploredge interface. On the left, a search bar contains the text 'catégorie de déchets'. Below it, a list of related terms is displayed, including 'boue d'égout', 'chiffon', 'composition des déchets', 'déchet agricole', 'déchet animal', 'déchet chimique', 'déchet d'action de dépollution', 'déchet d'emballage', and 'déchet de bois'. On the right, search results from Exalead are shown, including 'Andra, agence nationale pour la gestion des déchets radioactifs' and 'CPF rév. 1, 2003 - Catégorie - Déchets et débris de métaux...'. Annotations in red and blue highlight the search bar and the list of related terms, respectively.

Ce type d'outils se conçoit dans des dispositifs de veille ou pour certaines catégories d'utilisateurs ayant à réaliser de nombreuses recherches sur des domaines pas toujours bien maîtrisés. Le thésaurus constitue une carte d'orientation.

⁷ Voir aussi Fig.5. Aquabrowser Library

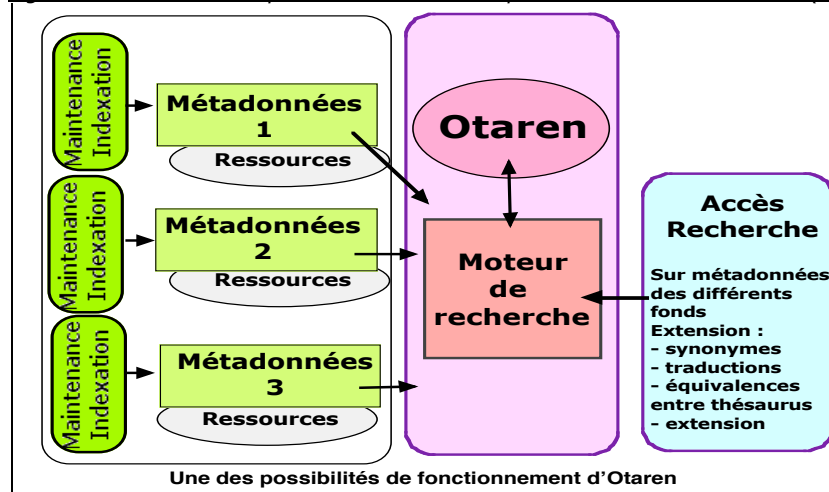
3.3. Un thésaurus pour améliorer la recherche fédérée

La recherche fédérée⁸ proposée par la plupart des portails documentaires nécessite le développement de connecteurs pour générer une requête par ressource interrogée. Ces connecteurs, délicats à réaliser, occultent bien souvent la question des vocabulaires, et de fait rendent la recherche aléatoire. C'est de notre point de vue la faiblesse d'un certain nombre de dispositifs actuels qui, s'ils offrent effectivement l'accès à de nombreuses ressources, le font au prix d'une baisse de la qualité de la réponse : beaucoup de bruit et de silence, avec peu d'intelligence en aval pour organiser le lot résultat et aider l'utilisateur à trouver ce qui lui convient.

Deux types d'actions sont possibles : l'une au niveau des traitements, l'autre en aval au niveau de la restitution des résultats.

- Si l'on reste dans des environnements de types bases de références bibliographiques (avec ou sans ressources numériques), l'accès fédéré pose l'épineux problème de la mise en correspondance entre les langages qui ont servi à indexer les documents. C'est dans ce contexte que le CNDP a développé **Otaren**⁹ (Fig.3). Cet outil est construit autour d'équivalences conceptuelles entre plusieurs langages contrôlés. Le prototype actuellement réalisé assure les équivalences entre : Motbis (le thésaurus de l'Education nationale), Rameau de la BNF ainsi que les thésaurus de l'ONU (UN-BIS) et de l'Unesco. Ce modèle permet d'exploiter les fonds existants sans avoir à passer par une réindexation des références. Nous renvoyons les lecteurs à l'article paru dans Documentaliste en 2007¹⁰. Cette même approche est mise en œuvre dans UMLS (Unified Medical Language System) qui « intègre près de deux millions de termes (dont 22 500 en langue française) provenant d'une cinquantaine de familles de thésaurus dans le domaine biomédical »¹¹ et de « nombreux projets de recherche et de développement recourent à l'UMLS ».

Fig.3 – Fonctionnement possible de tables d'équivalences entre thésaurus (Otaren)



- Lorsque l'on sort de ce cadre d'usage précis (bibliographique), que la nature des objets documentaires ou que les systèmes développés sont plus éloignés de ces contextes professionnels, lorsque le document numérique prend le pas sur sa référence, ou simplement lorsque l'on souhaite proposer un mode d'accès plus proches des questions des utilisateurs, d'autres technologies sont utilisables.

⁸ La recherche fédérée, Frédéric Cavazza (fichier PDF - 1,4 Mo), Présentation donnée en décembre 2005 lors d'un séminaire sur les intranet de nouvelle génération. http://www.fredcavazza.net/doc/Portail-Intranet-3G_Recherche-federee.ppt.pdf

⁹ Outil Thématique d'Aide à la Recherche pour l'Education Nationale.

¹⁰ « De la compatibilité à l'interopérabilité en matière de repérage d'information pertinente. La problématique et l'exemple d'Otaren, François Feyler, Documentalistes – Sciences de l'information 2007, vol..33, n°1, p.84-92 »

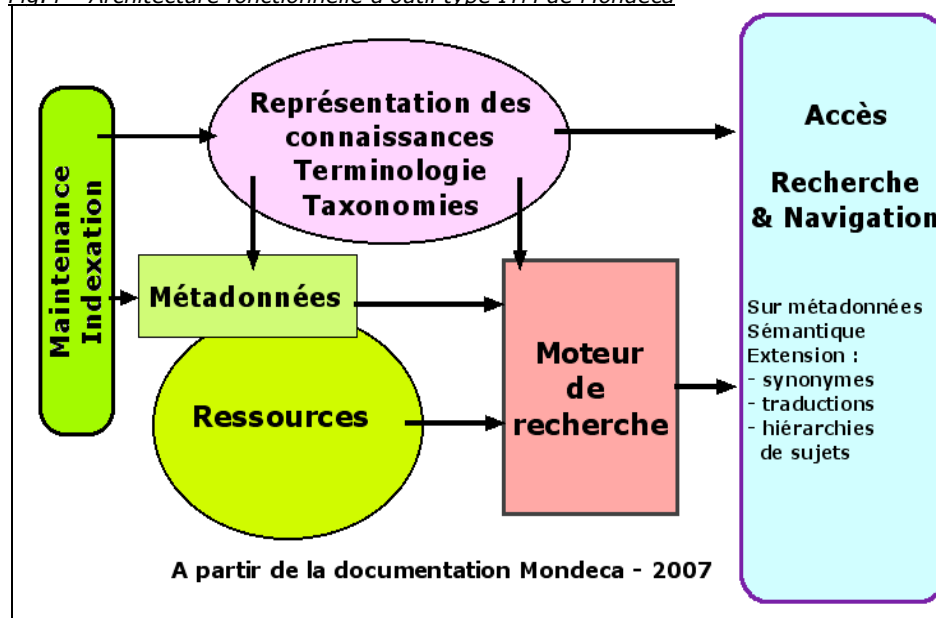
¹¹ Fiche d'information sur UMLS par Artist, 2005. http://artist.inist.fr/article.php3?id_article=206

Citons ITM (Fig.4) la solution proposée par la société française Mondeca (www.mondeca.com), « plateforme logicielle pour la gestion de référentiels métier, taxonomies, thésaurus, ontologies, bases de liens, centres de ressources, portails sémantiques ». Un travail préalable de modélisation est réalisé pour orienter les points d'accès par rapport aux « métiers » (tourisme, édition,...) et aux facettes spécifiques à ceux-ci. Le système, à la fois terminologique et de représentation structurée de la connaissance, est organisé selon ces « univers métier ».

Quelles sont les spécificités de ce type d'architecture :

- Les thésaurus de descripteurs et les index constitués à partir de ces thésaurus sont ré-utilisables dans ces applications¹².
- L'outil de représentation est autonome par rapport aux autres outils informatiques à savoir les outils de stockage des contenus ainsi que les moteurs de recherche.
- L'outil de représentation - schéma, ontologie, terminologie, peut être utilisé pour améliorer l'exploitation des ressources aussi bien grâce à des outils de fouille de données (Temis,...) ou des moteurs d'indexation plus traditionnels (Antidot, Exalead, Fast, Verity,...), utilisés conjointement.

Fig.4 – Architecture fonctionnelle d'outil type ITM de Mondeca



#4 - Présentation des résultats

Une fois la requête traitée par le moteur enrichi de ressources terminologiques, il est possible soit d'affiner la réponse, soit au contraire de l'étendre à partir d'une présentation structurée des résultats. Cette présentation structurée des résultats se fait par rapport à des métadonnées « métiers » adaptées aux utilisateurs ou plus riches sur le plan informatif que celles qui sont traditionnellement proposées grâce aux descripteurs ou plus simplement les thèmes ou classes du plan de classement auxquels ces références sont associées.

- Le site du Sénat propose d'affiner les résultats d'une recherche par rapport à une liste de thèmes¹³
- Dans le travail réalisé par F.Papy et S.Chauvin sur le catalogue de la bibliothèque universitaire (<http://visualcatalogue.univ-paris8.fr/>) le système exploite les parties des vedettes Rameau ainsi que les thèmes de la classification. Il offre de plus une représentation graphique associée à l'espace

¹² Des études montrent qu'il est dans certains cas difficile d'exploiter un thésaurus dans ces contextes. Voir : D'un thésaurus vers une ontologie de domaine pour l'exploration d'un corpus, Claude Chrisment, Nathalie Hernandez, Françoise Genova, Josiane Mothe, In Revue @metist, n°0, 2007, <http://dilib.inist.fr/~ducloy/AMETIST/public/0/IRIT/>

¹³ Chaque document est indexé par un ou plusieurs de ces thèmes. http://www.senat.fr/recherche/aide/aide_affinage.html

géographique des ressources documentaires.

- Le produit AquaBrowser (Fig.5) qui rencontre un franc succès dans de nombreux pays (sauf en France ?), propose un outil de navigation et/ou rebond à partir du lot résultat : un graphe (frame à gauche de l'écran) articule des éléments du lot résultat (relation d'association en blanc) pour affiner la recherche, ainsi que des termes issus d'un thésaurus de synonymes et d'équivalents linguistiques (traduction) pour étendre la recherche.

Fig.5 – Aquabrowser Library

Aquabrowser Library (<http://www.medialab.nl/>)

Un graphe de termes issus des contenus des notices (associations en blanc), ou d'un thésaurus d'équivalents (spelling variations et translations)

Des filtres selon les métadonnées dont celle sur les Sujets^{UP}

The screenshot shows the Aquabrowser Library interface. On the left, there is a term graph with 'india' at the center, connected to various related terms like 'buddhist', 'travel', 'Mandia', 'tibet', 'agra', 'east', 'mahal', 'indic', 'taj', 'passage', 'japan', 'china', 'pakistan', 'in-day', 'inda', 'indio', 'civilization', 'gandhi', 'south', 'india', 'british', 'whites', 'associations', 'yellow', 'spelling variations', 'green', 'translations', and 'blue', 'history trail'. The main search results area shows three results for 'India', including pilot guides, a DVD, and a book. A 'REFINE' sidebar on the right allows filtering by format, author, subject, and language. The URL <http://www.acornweb.org/> is visible at the bottom left.

En conclusion

A partir de ces quelques exemples que pouvons-nous dire du thésaurus à la recherche ?

Ce thésaurus à la recherche est :

- parfois plus étroit : uniquement les synonymes, ou bien quelques listes considérées comme structurantes comme point d'accès
- parfois plus étendu ; il intègre d'autres métadonnées que celles qui composent habituellement les thésaurus de descripteurs (le pays d'origine de l'auteur pour une recherche portant sur un nom de pays)
- parfois très extensif : avec des équivalents entre plusieurs thésaurus ou une intégration de sujets mis en équivalence des concepts (pont en béton armé équivalent de pont pour tel contexte d'usage)
- utilisé en back office à travers les possibilités d'expansions sémantiques (ITM de Mondeca...)
- exploitation des contenus des champs Descripteurs des notices, ou exploitation de mots-clés récupérés par les moteurs d'indexation mis en correspondance avec un thésaurus
-

Il semble possible de construire des interfaces de recherche riches apportant une substantielle aide à l'utilisateur en s'appuyant sur les thésaurus ou leurs cousins.

La complexité de ces dispositifs porte pour une grande part sur la prise en compte de l'environnement (langages déjà utilisés, type de ressources, pratiques des utilisateurs). Cela suppose une vision ouverte sur les technologies dont l'alliance avec les vocabulaires contrôlés reste une des solutions les plus prometteuses.

Annexe – Rapport entre concepts, sujets et classes

Dans certains contextes spécialisés, il serait possible d'envisager un thésaurus de descripteurs plus précis, plus spécialisés (par exemple : fabrication de plats cuisinés surgelés), qui autoriserait, soit la navigation dans cette taxonomie, soit une exploitation de celle-ci dans les résultats de recherche. Une autre solution est de construire une liste de « sujets » correspondant à des profils construits sur la base des descripteurs.

Fig.6 – Concepts – Sujets - Classe

Rôle des additifs dans la fabrication de plats cuisinés surgelés

