# Le problème de l'accès à l'information

Le problème de l'accès à l'information n'est pas neuf. Il a déjà été abordé dans le domaine des sciences documentaires, pour des collections de *documents-papier* dans un premier temps, pour des ensembles de ressources électroniques ensuite. Avec l'avènement du réseau Internet et du Web[1], c'est un nouveau type de collection documentaire qui est apparu. Son importance en ce qui concerne le nombre de documents et d'utilisateurs ainsi que l'accès largement public, au contraire de certaines archives présentes dans les entreprises et autres grandes organisations, ont alors entraîné une concentration importante des innovations dans ce secteur. Si le web reste un cas particulier de collection de documents, les technologies développées pour y accéder sont néanmoins souvent applicables d'une manière générale à tout ensemble documentaire numérique.

Actuellement, l'accès aux collections électroniques de documents est souvent réalisé à l'aide de mots clés. Ce système, s'il rencontre un certain succès, que ce soit sur le Web ou dans le cadre d'autres fonds documentaires, est loin d'être idéal. Le problème de l'ambiguïté lexicale et celui représenté par les multiples possibilités d'expression d'une information sont des obstacles importants au bon fonctionnement des systèmes de recherche. En fait, ces derniers maîtrisent difficilement tout ce qui fait la diversité et la richesse d'une langue naturelle. Une méthode de recherche performante se doit de prendre ces aspects en compte, voire même de les dépasser. Afin de maximiser la couverture et la précision d'une recherche par rapport à une collection de documents, il peut être profitable de passer d'un espace de mots à un espace de concepts. L'accès aux documents devrait donc idéalement se dérouler sur une base sémantique et non lexicale. Si cet objectif est assez ambitieux et encore en grande partie hors de portée des technologies actuelles, il n'en demeure pas moins intéressant de se demander comment, dans un premier temps, apporter des éléments de sens à la représentation et à l'indexation des documents. Ce qui rend cette tâche difficile, c'est le caractère souvent hétérogène des collections de documents qui entraîne de nombreuses difficultés lors de l'inventaire, de la manipulation, du jugement de la qualité et de la pertinence, et finalement de l'indexation même des documents.

[1] World Wide Web, désigne l'ensemble des documents disponibles sur le réseau Internet, reliés par des liens hypertextes et visualisables à l'aide d'un navigateur. Internet est le réseau informatique par lequel sont accessibles ces documents. L'usage courant confond souvent, de manière erronée, les deux termes.

D'abord, les ensembles de documents ne sont pas nécessairement organisés selon un plan précis, que ce soit logiquement ou physiquement. Ensuite, il existe une grande variété de formats de documents (formats de fichiers et organisation du texte dans le document), et leur contenu n'a pas toujours fait l'objet d'une validation. De plus, ces documents sont parfois difficilement accessibles[2]. Ces différents obstacles ne se retrouvent pas dans toutes les collections de documents, mais le Web en concentre une bonne partie. Certaines de ces difficultés étaient déjà connues et présentes avant l'expansion numérique, mais cette dernière a généralement eu un effet amplificateur, et les a rendues plus critiques. Concrètement, pour une ressource documentaire telle que le Web, un certain nombre de difficultés peuvent être mises en évidence :

- Le nombre de documents à traiter est tel qu'en pratique il est très difficile d'atteindre l'exhaustivité.
- La diversité thématique est très élevée, de nombreux domaines étant abordés.
- Le degré d'intérêt[3] des documents est variable. L'information proposée peut être cruciale ou très importante, ou au contraire complètement anecdotique.
- La qualité du contenu peut varier très fortement (la facilité de production et de diffusion permet à tout un chacun de produire des documents, indépendamment de toute contrainte éditoriale).
- L'authenticité des documents n'est pas toujours garantie et est parfois difficile à établir (possibilité de faux, difficulté de distinguer ce qui relève de l'opinion ou des faits, etc.).
- L'existence de redondances complètes suite à la diffusion par différents canaux, ou partielles suite à l'achat ou à la citation de contenu, opérations durant lesquelles le texte peut éventuellement être modifié.
- Les modes de diffusions numériques favorisent la circulation de documents parfois très courts qui ne présentent souvent que des informations partielles (par exemple les flux RSS, le système Twitter, etc.).
- L'information est disséminée en de nombreux endroits.
- L'existence d'une multitude de formats (encodage des caractères, format du document, structuration de l'information à l'intérieur du document, etc.).
- L'information est exprimée au moyen de beaucoup de langues différentes.

Face à ces obstacles, plusieurs domaines de recherche, principalement la *recherche d'informations* et l'*extraction d'informations* ont tenté de proposer des technologies pour améliorer l'accès à l'information selon des approches différentes. Après avoir brièvement introduit ces deux domaines à la section 1.2, nous proposerons un aperçu des différentes solutions qui ont été créées pour l'accès au Web à la section 1.3.

[2] Par exemple, les documents qui font partie de ce qui est appelé le *web invisible* n'ont pas de *pointeurs* permettant d'y accéder facilement.

[3] Cette notion est cependant en partie subjective. Ce qui est souligné ici est le fait que toutes les informations n'ont pas nécessairement le même statut.

## 1.2 Recherche d'informations et extraction d'informations

Grishman [1997] définit l'extraction d'informations (EI) comme étant :

> « the identication of instances of a particular class of events or relationships in a natural language text, and the extraction of the relevant arguments of the event or relationship. Information extraction therefore involves the creation of a structured representation (such as a data base) of selected information drawn from the text. »

Cette définition se situe dans la droite ligne de l'approche adoptée au cours des conférences MUC[4], *Message Understanding Conference* (Grishman et Sundheim [1996]), qui à partir du début des années 1990, ont contribué à fonder ce courant de recherche. Il peut sembler un peu réducteur de ne mentionner comme objet de l'extraction que les seuls événements et relations, mais ceux-ci peuvent être considérés selon une interprétation large qui se référera à un ensemble beaucoup plus vaste de types d'informations. D'aucuns préféreront cependant une formulation un peu plus générale, comme celle donnée par Moens [2006] :

> « Information extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, making the information more suitable for information processing tasks. » (p. 4)

L'extraction d'informations consiste donc à rechercher des éléments spécifiques, définis par la tâche d'extraction, dans des textes non structurés (en langage naturel) et à les caractériser selon les catégories définies au préalable. Ce processus peut-être vu comme une étape de (pré)traitement destiné à produire un document plus propice au traitement automatique, ou au contraire, si les informations extraites constituent le résultat attendu, comme un aboutissement.

En recherche d'informations (RI), l'approche est différente. Baeza-Yates et Ribeiro-Neto [1999] en exposent le principe général :

> « the primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few as non-relevant documents as possible. » (p. 2)

Un aspect important réside dans l'ordre de présentation des résultats :

> « To be effective in its attempt to satisfy the user information need, the IR system must somehow 'interpret' the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query.» (Baeza-Yates et Ribeiro-Neto [1999], p. 2)

L'activité de recherche implique une tâche préalable : l'*indexation* des documents. Celle-ci peut être effectuée selon diverses méthodes et produire différents types d'index. La recherche d'informations se déroule donc la plupart du temps en deux phases. Tout d'abord, les documents sont analysés afin d'y relier des clés d'indexation ou de les classer dans des catégories. Ensuite, la recherche consiste

[4] `http://www-nlpir.nist.gov/related_projects/muc/index.html`

à comparer les requêtes formulées par les utilisateurs à cet index afin de retrouver les documents pertinents.

La distinction faite entre extraction d'informations et recherche d'informations n'est, dans la pratique, pas si tranchée. En effet, l'extraction peut faire appel à des techniques de recherche, et inversement. Par exemple, les systèmes de classification mis au points en RI peuvent être utilisés en amont de l'EI afin de séparer les documents en sous-corpus plus homogènes ou, de manière encore plus fine, pour sélectionner des phrases à analyser de manière plus détaillée (Nédellec *et al.* [2001]). De même, l'EI peut, entre autres, réduire un document représenté initialement par son contenu entier à un ensemble particulier de mots ou d'expressions et ainsi diriger l'indexation (Riloff et Lehnert [1994], Fairon et Watrin [2003]). Les deux domaines sont donc complémentaires.

Dans cette thèse, nous nous intéressons principalement à la recherche d'informations, en tant que moyen d'améliorer l'accès aux documents et, par conséquent, à l'information qu'ils contiennent. L'extraction d'informations sera cependant massivement utilisée pour atteindre cet objectif. Plus particulièrement, l'analyse temporelle présentée à la partie II, relève de l'EI mais est finalement mise au service du système de classification et d'indexation présenté au chapitre 8. Bien entendu, ce système ne représente qu'un exemple possible d'utilisation de l'analyse temporelle parmi bien d'autres. Les développements consentis en la matière sont donc exploitables de diverses manières, que ce soit pour des applications en recherche ou en extraction d'informations.

## 1.3 Les systèmes de recherche d'informations

Avant toute chose, précisons que nous écartons de la recherche d'informations, les systèmes purement encyclopédiques, telles que Universalis[5] ou Wikipedia[6]. Même si ceux-ci satisfont à un certain nombre de critères que nous attendons d'un système de recherche d'informations performant, c'est-à-dire, entre autres, un accès à l'information partiellement basé sur le sens (grâce à des classification par catégories ou par thèmes) ou une certaine qualité de l'information[7], ils doivent avant tout être considérés comme un ensemble de documents parmi d'autres. En effet, la couverture thématique et surtout la diversité des documents proposés est forcément limitée. Nous nous intéressons ici, au contraire, aux méthodes rendant possible l'accès à une collection quelconque de documents (textuels) numériques, potentiellement très vaste, dont l'exemple le plus parlant est le Web[8].

Les obstacles présentés à la section 1.1 expliquent en grande partie pourquoi, encore aujourd'hui, il est parfois ardu de trouver les documents pertinents parmi ce type de ressources, et ce malgré les efforts pour développer des systèmes de recherche performants. Comme nous l'avons déjà mentionné, ces systèmes procèdent généralement en deux temps : l'indexation des documents, qui permet en-

[5] `http://www.universalis.fr`

[6] `http://www.wikipedia.org`

[7] La question de la qualité de l'information fournie par une encyclopédie de type collaboratif, telle que Wikipedia, peut être discutée, mais sort de notre propos.

[8] Dans les paragraphes qui suivent, nous citons à de nombreuses reprises des exemples issus du Web. Celui-ci concernant à la fois un très grand nombre de documents et d'utilisateurs, les technologies de recherche d'informations se sont naturellement développées dans ce milieu. Les principes évoqués restent cependant valables dans le cadre d'un intranet ou d'une collection privée de documents électroniques.

suite une interrogation de l'index au moyen d'une requête. Cette dernière étape se décompose plus précisément en deux parties : d'une part, la formulation de la requête, et d'autre part, la confrontation de celle-ci à l'index.

Au cours du temps, les techniques d'indexation ont bien entendu évolué dans le but de concilier deux objectifs *a priori* opposés, l'efficacité du processus de traitement et l'obtention d'une représentation la plus complète et la plus adéquate possible du document dans l'index. L'indexation peut être réalisée de manière manuelle ou automatique et les clés d'index peuvent se situer dans l'espace des mots, sous la forme de mots clés librement choisis, ou dans un espace plus conceptuel, dont les éléments – des catégories – sont prédéfinis et porteurs d'un sens précis.

En ce qui concerne les requêtes, il existe divers moyen de les exprimer : à l'aide de mots clés, en utilisant le langage naturel, par sélection ou navigation dans un ensemble de catégories prédéfinies ou encore en passant par une architecture de facettes.

Quant au processus de confrontation de la requête à l'index, il peut faire intervenir divers processus, automatiques ou requérant l'intervention de l'utilisateur.

Ces aspects sont exposés et illustrés au fur et à mesure de l'examen des différents systèmes. Après un rapide aperçu des premiers développements (Section 1.3.1), les systèmes actuels sont passés en revue selon qu'ils utilisent un espace fermé ou ouvert de clés d'indexation (Sections 1.3.2 et 1.3.3). Finalement, les dernières évolutions en matière de moteurs sémantiques sont abordés (Section 1.3.4).

### 1.3.1 Les premiers systèmes

Au début des années 1990, les premiers développements[9] permettant de rechercher des documents sur Internet furent à l'image du réseau auquel ils s'appliquaient, c'est-à-dire limités, surtout en comparaison avec ce qui a vu le jour par la suite. Le premier moteur de recherche, *Archie*[10], se résumait à une simple liste de documents qui permettait des requêtes sur les noms de ceux-ci. Par la suite, un moteur tel que *JumpStation*[11] a permis d'étendre la recherche à une partie limitée du document, les titres en l'occurrence. Avec l'intensification de l'utilisation du réseau Internet et du Web, la quantité de documents devint ensuite de plus en plus importante. Les outils de recherche d'informations s'adaptèrent et se développèrent alors en conséquence, pour finalement aboutir aux systèmes que nous connaissons aujourd'hui. Ceux-ci sont présentés au cours des sections suivantes.

[9] Une présentation de l'histoire des moteurs de recherche peut être consultée sur le site `http://www.searchenginehistory.com`.

[10] Créé à la McGill University, à Montréal, en 1990.

[11] Lancé en 1993 à la University of Stirling, Écosse.

### 1.3.2 Indexation dans un espace fermé de clés

#### *Les répertoires de liens*

Les répertoires de liens construits de manière manuelle sont probablement parmi les systèmes les plus simples technologiquement parlant, mais ils ont malgré tout atteint une certaine popularité, surtout au début du Web. Citons par exemple l'*Open Directory Project* (dmoz)[12] (Figure 1.1), qui continue d'ailleurs encore aujourd'hui de proposer un répertoire entretenu et enrichi manuellement par des éditeurs humains bénévoles. Chaque site web intégré dans la ressource est ajouté à la catégorie qui lui correspond le mieux. L'ensemble des références est organisé selon une hiérarchie de catégories parmi lesquelles l'utilisateur navigue afin de trouver la section qui l'intéresse, pour ensuite explorer les liens qui y sont contenus. Alternativement, une recherche par mots-clés (voir section 1.3.3), limitée aux sites référencés dans le répertoire, est également proposée.


*Figure 1.1 : Le répertoire de liens Open Directory Project (dmoz), construit manuellement par des éditeurs bénévoles.*

#### *L'utilisation de terminologies*

Le principe de recherche par catégories se place dans la droite ligne des répertoires, tout en offrant un peu plus de souplesse et de puissance, en permettant par exemple d'indexer un document à l'aide de plusieurs catégories. Celles-ci peuvent être organisées de manière plus ou moins complexe : liste de termes, taxonomie, thésaurus voir même ontologie. La définition de ces catégories étant une tâche compliquée, et donc longue et coûteuse, leur structuration n'atteint cependant pas toujours les formes

[12] `http://www.dmoz.org`. Fondé en 1998, en réaction à l'attitude jugée trop peu réactive de Yahoo (`http://www.yahoo.com`), qui proposait également un répertoire de liens depuis 1994. Yahoo a ensuite évolué vers un modèle d'indexation automatique et d'interrogation par mots-clés libres, qui est actuellement le plus courant.

les plus complexes. Une fois la terminologie[13] ou la classification disponible, les documents peuvent y être indexés. Généralement, c'est un documentaliste expert du domaine qui attribue manuellement une ou plusieurs catégories à chacun d'entre eux. Si cette méthode se révèle à nouveau assez onéreuse, elle garantit cependant une indexation d'une qualité assez élevée, et qui est effectuée dans l'espace des concepts et non pas celui des mots (Chaumier et Dejean [2003], Da Sylva [2004], Da Sylva [2006]). En raison de son coût, cette méthode est plutôt utilisée par les entreprises ou grandes organisations, mais peut aussi être appliquée pour de plus petites collections de documents.

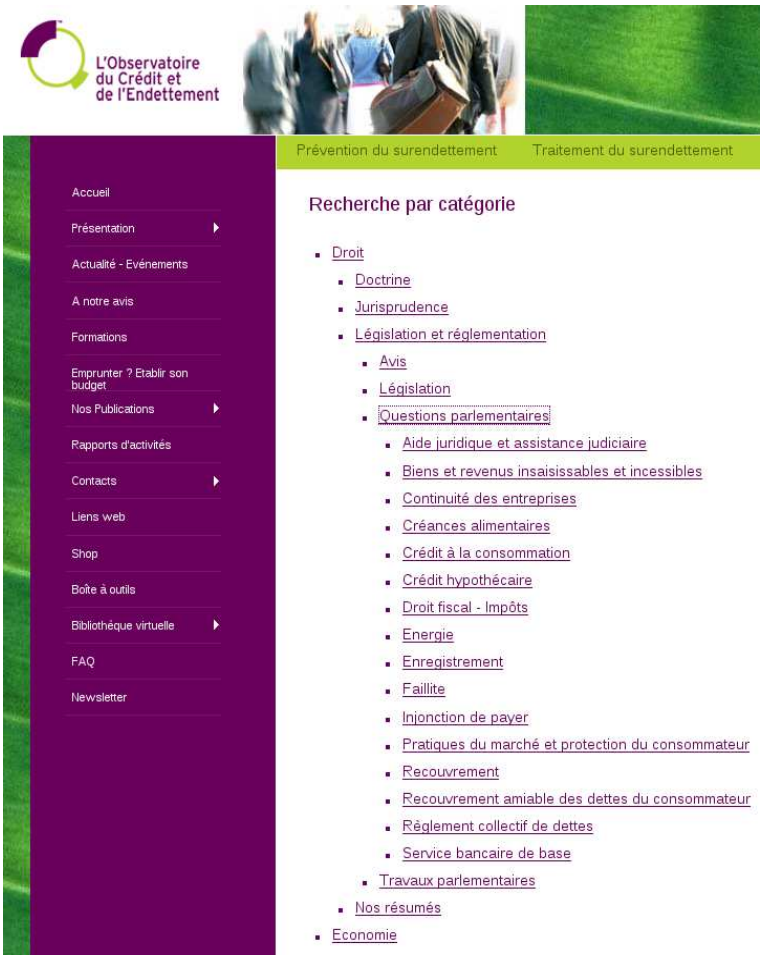


*Figure 1.2 : Exemple de recherche par catégories prédéfinies (Site : Observatoire du Crédit et de l'Endettement).*

Pour la recherche dans ce type de système (Figure 1.2[14]), l'utilisateur se voit proposer un certain nombre de clés prédéfinies parmi lesquelles il doit faire son choix. Ces catégories peuvent être présentées de manière plate ou hiérarchique. La sélection d'une, ou éventuellement de plusieurs catégories entraîne l'affichage des documents s'étant vus attribuer au moins une de ces catégories lors de la phase d'indexation. Ce mode de recherche peut aussi éventuellement être combiné à une requête par

[13] Le sens donné au mot *terminologie* est ici celui qui désigne une ressource, telle que celles exposées à la section 1.4.1, qui reprennent un ensemble de termes relatifs à un ou plusieurs domaines, activités, etc.

[14] `http://www.observatoire-credit.be` (consulté le 31/07/2010). Voir également le site du Sénat dont la recherche avancée (`http://www.senat.be/www/?MIval=/index_senate&MENUID=12420&LANG=fr`) propose un accès aux documents au travers des catégories du thésaurus Eurovoc (`http://eurovoc.europa.eu`).

mots clés. C'est par exemple le cas de l'interface expérimentale de JSTOR[15], qui permet de chercher des référence bibliographiques par mots-clés (voir section 1.3.3), et d'ensuite réordonner le résultat en pondérant l'importance des différentes catégories (thèmes) concernées par les résultats de la requête (Figure 1.3).

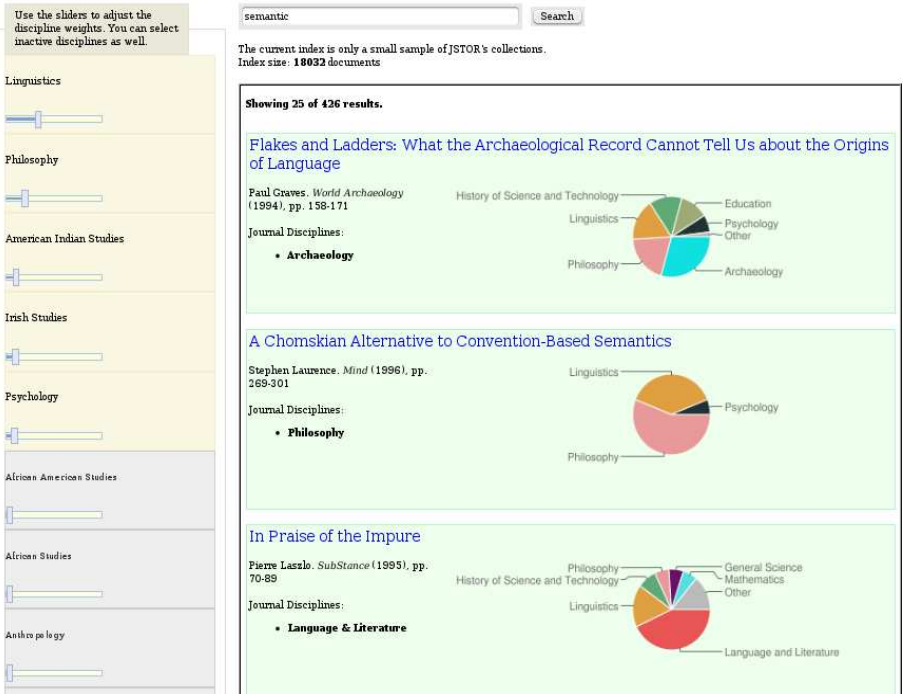


*Figure 1.3 : Exemple de recherche par mots-clés et catégories.*

## *La recherche par facettes*

Une variation de ce mode d'interrogation est la recherche par facettes. Basé sur le travail initial de Ranganathan [1967], la pertinence de l'application de ce principe dans le cadre des ressources électroniques modernes, telles que le Web, a été démontrée (Zins [2002], Kyung-Sun *et al.* [2006]). Avec la recherche, ou navigation, par facettes, les catégories sont regroupées en plusieurs groupes (les facettes) représentant chacun une caractéristique particulière des documents. L'utilisateur est invité à utiliser plusieurs de celles-ci, de manière simultanée ou successive, afin de raffiner le résultat proposé par le système. Ce type de recherche présente l'avantage de pouvoir plus facilement caractériser des objets complexes, dont la nature peut être définie selon plusieurs axes. Par exemple, le portail *Innovons*[16] (Figure 1.4) a pour vocation d'indexer des documents relatifs à l'*innovation*, sujet qui touche potentiellement à tous les domaines scientifiques et techniques (facette 2), et qui peut s'appliquer à divers secteurs d'activités (facette 3), dans le but de fournir différents types de produits (facette 4), au travers d'un ensemble de métiers et compétences (facette 5). Dans ce cadre, le portail propose d'apporter une réponse à une série de besoins concrets (facette 1).

[15] `http://dbrowser.jstor.org/browser.cgi?q=semantic&btnG=Search`, accédé le 02/12/2010.
[16] `http://www.innovons.be`

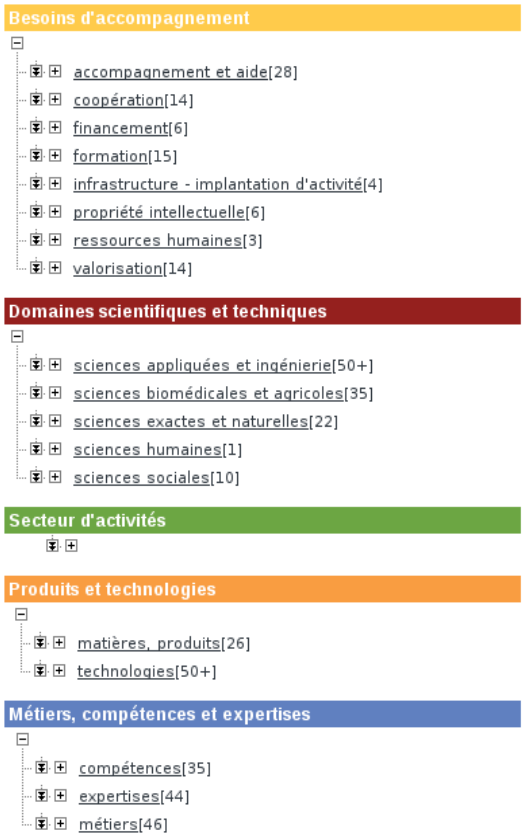


*Figure 1.4 : Exemple de catégories organisées par facettes (Site : Innovons en Région wallonne).*

### 1.3.3 Indexation dans un espace ouvert de clés

#### *Recherche par mots clés*

Les moteurs de recherche, dont les principaux représentants[17] sont aujourd'hui Google, Yahoo, Baidu et Bing (Microsoft), fonctionnent selon une autre philosophie. Leur principe est d'entretenir, de manière automatique, un index qui exploite principalement le contenu même du document ainsi que certaines métadonnées. Cet index est ensuite confronté aux requêtes formulées sous la forme de mots clés afin de fournir la liste des documents jugés les plus pertinents (Figure 1.5[18]). L'utilisation de ce mode d'interrogation laisse une grande liberté à l'utilisateur : il peut choisir les termes exacts de sa requête et ne doit pas respecter un formalisme spécifique. Il est cependant souvent possible d'utiliser des guillemets pour délimiter des expressions composées, ainsi que certains opérateurs logiques (AND, OR, NOT). En pratique, rares sont les utilisateurs à employer réellement ces possibilités[19].

[17] Selon une statistique établie par Comscore et relayée par le Journal du Net (`http://www.journaldunet.com/cc/03_internetmonde/intermonde_moteurs.shtml`, consulté le 31/07/2010), Google obtiendrait 67,5% des parts de marché au niveau mondial en juillet 2009, contre 7,8% pour Yahoo, 7% pour le moteur chinois Baidu et 2,9% pour Bing.

[18] `http://www.google.be`

[19] L'observation du nombre de mots inclus dans les requêtes montre que celui-ci est assez faible : Assadi et Beaudouin [2002] établissent que trois-quarts des requêtes ont une longueur inférieure ou égale à deux mots (les dernières tendances montrent cependant un allongement progressif des requêtes, selon un rapport Hitwise : `http://image.exct.net/lib/fefc1774726706/d/1/SearchEngines_Jan09.pdf`). Cette longueur peu importante n'invite évidemment pas à l'utilisation d'opérateurs complexes. Jansen et Eastman [2003] rapportent que environ 10% des requêtes utilisent des opérateurs booléens (13% pour Assadi et Beaudouin [2002]), et établissent que leur apport est sou-


*Figure 1.5 : Page de résultats fournis par Google suite à une recherche par mots clés.*

Dans ce type de système, la qualité du résultat est aussi grandement influencée par l'ordre de présentation des documents. Le moteur se doit, dans la mesure du possible, de proposer en premier lieu le document le plus pertinent, et de continuer ensuite par ordre décroissant d'importance. Par exemple, l'algorithme Page Rank (Brin et Page [1998]) utilisé par Google à cet effet, permet d'exploiter les interconnexions entre les documents afin de faire ressortir les pages les plus *importantes*.

L'avantage déterminant des moteurs de recherche est leur capacité à tenir à jour de manière automatique des index répertoriant un nombre très élevé de documents. Lorsque des liens existent entre ceux-ci, le système est capable de *découvrir* tout seul les nouveaux documents. Au delà de cet aspect, ils rencontrent néanmoins certaines difficultés à appréhender toutes les finesses du langage naturel. En plus des variations grammaticales (singulier/pluriel, forme nominale/forme adjectivale, etc.), qui sont maintenant gérées dans une certaine mesure, la principale difficulté est de pouvoir prendre en compte la nature intrinsèquement variée et ambiguë de la langue. En effet, d'une part, un concept peut souvent être désigné par de nombreux mots ou expressions composées, et d'autre part un mot peut également référer à plusieurs concepts. Cette relation *de plusieurs à plusieurs* entre l'espace des concepts et celui des mots explique pourquoi une requête classique à partir de mots clés ramène rarement l'ensemble des documents pertinents et, dans le même temps, propose souvent des résultats qui ne sont pas en rapport avec la recherche de l'utilisateur. En dépit de cette faiblesse, la technologie de recherche par mots clés est toujours aujourd'hui celle qui est la plus utilisée, aussi bien en entreprise que par le grand public.

### *Recherches avancées*

Les recherches dites *avancées* correspondent à une variation des recherches par mots clés pour lesquelles le système attend de la part de l'utilisateur des clés de recherche qui correspondent à des types d'information définis a priori. On retrouve assez fréquemment ce type d'interrogation sous la forme

vent très faible en ce qui concerne la qualité des résultats, à moins que l'utilisateur ait une connaissance assez pointue du fonctionnement du moteur de recherche.

d'un formulaire proposant des champs spécifiques pour les principales catégories de clés gérées par le système. Pour une bibliothèque, on se verra par exemple proposer un champ pour le nom de l'auteur, un autre pour le titre, un troisième pour la maison d'édition, et ainsi de suite. Cet exemple est parfaitement illustré par le formulaire de recherche proposé par Google Livres[20] (Figure 1.6). Ce type de requête s'applique de préférence à des documents au moins partiellement structurés, ou proposant les métadonnées nécessaires.

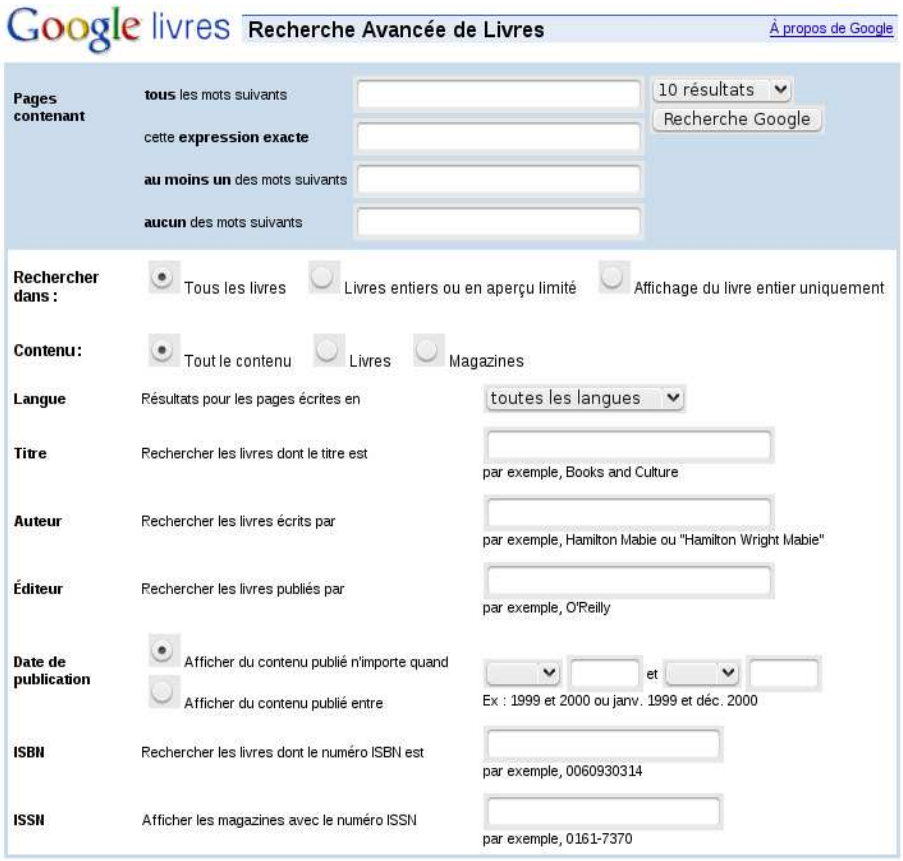


*Figure 1.6 : Formulaire de recherche avancée (Google Livres) permettant de contraindre la recherche sur certains types de données.*

### 1.3.4 Les moteurs *sémantiques*

Une nouvelle génération de moteurs, dits *sémantiques*, a fait son apparition depuis quelques années. D'une manière générale, on parle du *Web sémantique* (Berners-Lee et Lassila [2001]), dans le cadre duquel chaque document est accompagné d'un ensemble de données sémantiques qui décrivent son contenu. Cette couche supplémentaire doit permettre à un logiciel d'accéder directement au sens de l'information et non plus à sa matérialisation sous la forme de mots. Cela ouvre évidemment de nombreuses perspectives pour l'amélioration des résultats et l'augmentation de la complexité des recherches. Cependant, la difficulté que représente la production de données correctement annotées et leur exploitation est importante, et cela a pour conséquence que peu d'applications tirent déjà

[20] `http://books.google.be`

complètement parti de tous les aspects du Web sémantique. En pratique, de nombreux systèmes proposent certaines évolutions, qui ne s'inscrivent pas nécessairement dans ce cadre strict, mais qui permettent tout de même d'apporter des éléments de sens aux documents, comblant ainsi en partie le fossé entre l'espace des mots et l'espace des concepts. C'est par exemple le cas des techniques d'extension de requêtes.

### *Extension de requêtes*

La richesse des langues naturelles se traduit par une variété et une ambiguïté du lexique et de son utilisation. En recherche d'informations cela a pour conséquence qu'on observe souvent un écart important entre le contenu lexical des requêtes des utilisateurs et celui des documents (Cui *et al.* [2002]). Les recherches menées sur les possibilités d'extension de requêtes constituent, à cet égard, une démarche intéressante. Le terme *extension* est cependant quelque peu trompeur puisqu'il vise, en réalité, à la fois l'augmentation de la couverture et l'élimination des résultats non pertinents afin d'augmenter la précision. Les techniques utilisées sont assez nombreuses et variées, leur présentation dans le cadre de cette introduction sera donc rapide et forcément incomplète. Ces méthodes nécessitent aussi souvent la résolution de problèmes connexes tels que la désambiguïsation du sens des mots ou la prise en compte de l'aspect multilingue des documents (Gaillard *et al.* [2010]). Ces aspects ne peuvent être considérés comme des tâches triviales et représentent à eux seuls divers défis.

L'idée de l'extension de requête n'est pas neuve puisque Salton et Lesk [1968] montraient déjà que l'usage de synonymes pouvait améliorer les résultats des systèmes de recherche d'informations. Par la suite, des ressources telles que WordNet ont souvent été utilisées afin d'étendre les termes de requêtes avec des termes possédant un sens commun (Voorhees [1994], Moldovan et Mihalcea [2000]).

Des analyses plus complexes peuvent également permettre d'aller plus loin dans l'enrichissement des requêtes. L'exploitation d'informations issues d'ontologies peut par exemple venir compléter la recherche initiale (Bhogal *et al.* [2007], Guelfi *et al.* [2007]). Au delà des synonymes, que nous avons déjà mentionnés, différentes informations peuvent être extraites : hyperonymes, hyponymes, méronymes, ou encore d'autres relations sémantiques (Joho *et al.* [2002]).

L'analyse des logs de requêtes est aussi souvent utilisée pour trouver des termes de recherche associés à la requête initiale. Cui *et al.* [2002] utilisent une méthode consistant en l'extraction, à partir de ces logs, de corrélations probabilistes entre les termes des requêtes contenues dans les logs, et les termes provenant des documents. Les probabilités ainsi obtenues permettent alors d'ajouter les termes qui semblent les plus appropriés à l'extension d'une nouvelle requête.

Une autre idée largement exploitée est la technique de *relevance feedback* (Salton et McGill [1983]) dont le principe est d'utiliser les mots issus des documents les plus pertinents dans la liste de résultats originale. Cette méthode nécessite une participation relativement importante de la part de l'utilisateur. Afin de minimiser celle-ci, il est possible de sélectionner systématiquement les documents les mieux classés dans la liste de départ. Il est évidemment nécessaire que l'algorithme d'ordonnancement des documents donne de bons résultats dès le début, pour que cette approche puisse fonctionner.

Enfin, à l'aide des techniques de *clusering*, il est aussi possible de rassembler les résultats obtenus en plusieurs groupes (*clusters*) représentant diverses interprétations qui ont pu être distinguées (Stefanowski et Weiss [2003]). Cette approche a en particulier été implémentée pour le moteur de recherche Carrot2[21]. La figure 1.7 montre les différents clusters, et les sens correspondants, construits à partir de la recherche « extraction ». On y retrouve entre autres « information extraction », « DNA extraction », ou encore « tooth extraction », qui représentent effectivement des sens assez différents.

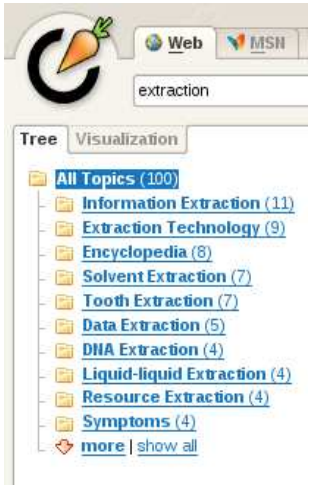


*Figure 1.7 : Le moteur de recherche Carrot2 organise les résultats en groupes correspondant aux interprétations qu'il a pu identifier, afin de préciser la recherche initiale (ici, « extraction »).*

Toutes ces techniques peuvent être utilisées de différentes manières. La requête initiale peut par exemple être étendue directement afin de présenter dès le départ un résultat *amélioré* à l'utilisateur. Le mécanisme peut aussi s'opérer en deux temps, en faisant intervenir la participation de l'utilisateur suite à l'introduction de sa requête. L'analyse de cette dernière, ou des résultats obtenus par son exécution, est dans ce cas utilisée pour fournir diverses possibilités d'affinement de la requête initiale ou pour orienter l'utilisateur vers de nouvelles recherches, sémantiquement proches.

À titre d'illustration, les propositions émises par le moteur de recherche de Yahoo (Figure 1.8), semblent faire intervenir plusieurs des techniques évoquées. À partir d'une requête « énergie », on remarque dans la première colonne, ce qui ressemble à des mots-clés assez proches (« énergie éolienne », « radio energie », *etc.*), et qui pourraient typiquement être obtenus par analyse de logs de requêtes. Dans les colonnes suivantes, on voit par contre apparaître des expressions plus éloignée au niveau lexical (« électricité ») mais aussi sémantique (« développement durable », « chaleur », « fournisseur de gaz »), ce qui suggère l'intervention de ressources ou procédés sémantiques plus complexes.

[21] `http://search.carrot2.org`


*Figure 1.8 : Le moteur de recherche Yahoo fournit des propositions d'affinement de requêtes ainsi que l'orientation vers des recherches sémantiquement proches.*

## *Le web sémantique*

La description sémantique du contenu des documents, telle que proposée dans le cadre du Web sémantique (Berners-Lee et Lassila [2001]), a pour objectif de permettre l'élaboration d'applications complexes. Celles-ci exploitent de manière automatique ces données structurées afin de fournir un service particulier, par exemple fournir l'adresse et les heures d'ouvertures d'un restaurant italien, dans une certaine ville, et qui propose un plat bien spécifique. Cela suppose évidemment des possibilités relativement étendues de recherche et d'analyse de l'information. Cette situation semble encore loin d'être atteinte à l'heure actuelle.

L'enrichissement des requêtes en données sémantiques par l'utilisateur lui-même pourrait cependant être considéré comme un premier pas dans cette direction. Cette approche, proposée par Umbrich et Blohm [2008], permet par exemple de préciser, à l'aide d'un balisage XML, que tel mot-clé correspond à un lieu, et que le résultat attendu est une personne. La recherche s'applique à une collection de documents qui a au préalable été indexée à l'aide de Wikipedia[22] et de l'ontologie YAGO[23] (Suchanek *et al.* [2007]). Shah *et al.* [2002] ont une approche relativement similaire, à la différence qu'un mécanisme de traduction transforme automatiquement une requête classique en requête sémantique, exprimée à l'aide du langage DAML+OIL.

La formulation de requêtes en langage naturel, proposée entre autres par Powerset[24], Hakia[25] ou encore Ask[26], suppose également un minimum de prise en compte d'éléments sémantiques. Bien que les résultats présentés soient effectivement différents de ceux obtenus avec un moteur classique (comme à la figure 1.9), il n'apparait pas toujours clairement si le moteur a réellement interprété la question ou s'il l'a juste transformée en un ensemble de mots-clés. D'autre part, il faut aussi souligner que la plupart des moteurs sémantiques se limitent à l'exploitation de ressources particulières, au moins partiellement structurées ou enrichies de données sémantiques, telles que Wikipedia.

Il ne faut pas non plus confondre ce type de moteur avec les systèmes de question-réponse, tels que True Knowledge[27] ou Wolfram[28], qui effectuent de véritables raisonnements et exploitent également des bases de connaissances internes afin de fournir des réponses et non une liste de documents.

[22] `http://www.wikipedia.org`
[23] `http://www.mpi-inf.mpg.de/yago-naga/yago/`
[24] `http://www.powerset.com`
[25] `http://www.hakia.com`
[26] `http://fr.ask.com/`
[27] `http://www.trueknowledge.com`
[28] `http://www.wolframalpha.com`

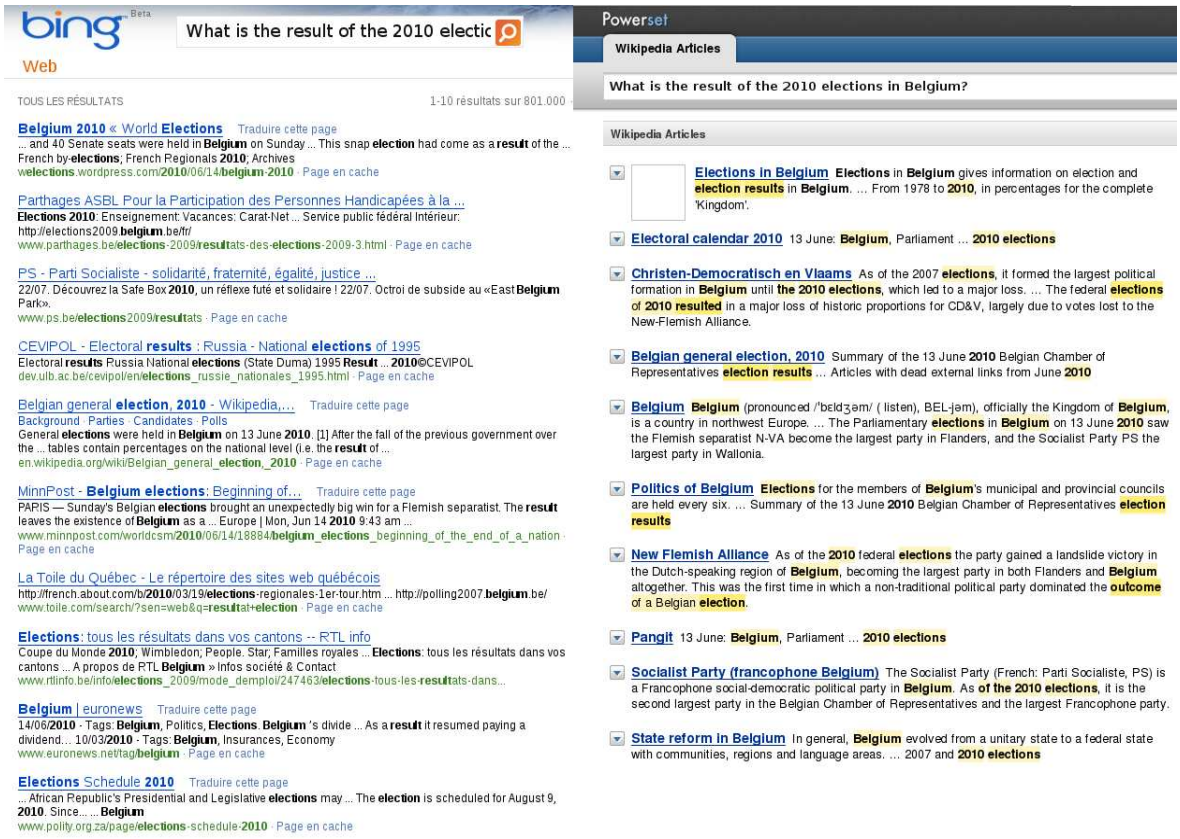


*Figure 1.9 : Comparaison des résultats entre le moteur de type mots-clés (bing) et le moteur sémantique (Powerset) de Microsoft.*

Au final, si les technologies *sémantiques* poursuivent un objectif légitime, elles semblent avoir quelques difficultés à s'imposer. De nombreux obstacles restent à surmonter en la matière, entre autres la production des documents aux formats et selon les standards définis (ce qui implique un effort important), la question de l'interopérabilité entre ontologies, etc.

## 1.4 Les systèmes de catégories en tant que couche sémantique pour la recherche d'informations

À l'issue de ce tour d'horizon des technologies de recherche d'informations, il semble clair que le couple indexation *full text* et recherche par *mots-clés*, même s'il reste le système le plus fréquemment utilisé, comporte de nombreuses limitations. Il faut cependant concéder que les espoirs qui ont pu être placés dans les technologies *sémantiques* n'ont apporté pour l'instant que peu de solutions concrètes et à grande échelle.

Comment dès lors dépasser les difficultés liées à la langue naturelle et apporter des éléments de sens lors de l'activité de recherche d'informations ? Un élément de réponse peut être trouvé dans l'utilisation d'un ensemble fermé de clés ou de catégories lors de l'indexation et de la recherche (voir section 1.3.2). Cette approche, si elle n'est pas sans poser certains problèmes (qui seront examinés à la section 1.4.3), permet effectivement d'avoir un certain contrôle sémantique sur l'indexation car celle-ci délaisse l'espace des mots pour s'effectuer au moyen de catégories qui possèdent par défi-

nition un sens particulier. La section 1.4.1 présente un certain nombre de *systèmes terminologiques* qui peuvent servir à définir un ensemble cohérent et organisé de catégories. Quelques exemples de systèmes qui proposent un mode de recherche par catégories sont ensuite passés en revue à la section 1.4.2. Finalement, pour conclure, la section 1.4.3 dresse un bilan des avantages et inconvénients de l'utilisation d'un ensemble fermé de clés, que ce soit pour pour l'indexation ou la recherche, avant de finir sur les perspectives offertes dans ce domaine.

### 1.4.1 Les systèmes terminologiques

Un ensemble fermé de clés, ou de catégories, nécessaire à l'indexation peut être organisé de diverses façons plus ou moins complexes à l'intérieur d'une terminologie[29].

#### *Vocabulaire contrôlé*

L'appellation *vocabulaire contrôlé* constitue la désignation générale de tout ensemble de termes, définis et sélectionnés par un ensemble d'experts. Un tel vocabulaire constitue donc un sous-ensemble du vocabulaire complet d'une langue (de plusieurs langues lorsqu'il s'agit d'une ressource multilingue). Il est généralement mis au point de manière à couvrir et à décrire un ou plusieurs domaines particuliers. Son utilité est de permettre l'organisation des connaissances à des fins de recherche d'informations.

Les termes qui constituent le vocabulaire contrôlé peuvent constituer une simple liste (par exemple le vocabulaire RAMEAU[30]) ou être organisés de diverses manières : taxonomie, thésaurus ou ontologie.

#### *Taxonomie*

La taxonomie est une forme assez simple de vocabulaire contrôlé. Elle consiste à organiser les termes à l'aide de relations hiérarchiques. Les taxonomies sont souvent utilisées dans le domaine des sciences de la nature, pour classer les différentes espèces animales et végétales. Par exemple, une taxonomie des virus a été mise au point par l'ICTV[31].

#### *Thésaurus*

Un thésaurus est un vocabulaire contrôlé un peu plus complexe. Les termes qu'il regroupe, et qui permettent de définir et de décrire les concepts d'un certain domaine utilisés par un groupe de personnes, sont liés de manière hiérarchique et transversale.

[29] Une rapide introduction peut être obtenue dans l'article de Woody Pidcock, « What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model ? » (`http://www.metamodel.com/article.php?story=20030115211223271`, date de dernière consultation : 13/08/2010).

[30] *Répertoire d'autorité-matière encyclopédique et alphabétique unifié*, langage documentaire élaboré et utilisé, entre autres, par la Bibliothèque nationale de France. `http://rameau.bnf.fr`

[31] *International Committee For Taxonomy Of Viruses*, `http://www.ictvonline.org/index.asp?bhcp=1`.

Un concept est représenté par un terme principal appelé *descripteur* qui peut être relié à plusieurs *non descripteurs* ou *synonymes* par une relation *used-for* (UF). Les concepts sont organisés hiérarchiquement à l'aide des relations *broader-than* (BT) et *narrower-than* (NT). La relation *related-term* (RT) permet de définir un lien de similarité entre deux concepts. Les grands thésaurus peuvent être fragmentés en *microthésaurus* qui couvrent chacun un sous-thème particulier. Plusieurs normes internationales dont ISO [1986] et AFNOR [1981] définissent plus précisément les thésaurus.

La portée d'un thésaurus peut être très large, par exemple Eurovoc[32], ou au contraire très spécialisée, tel que Agrovoc[33]. Ces deux thésaurus comptent un grand nombre de niveaux hiérarchiques et de descripteurs[34]. De nombreuses organisations se contentent cependant de vocabulaires de taille plus modeste. Van Slype [1987] préconise l'usage de 500 à 1.500 descripteurs pour des bases de données ayant un accroissement de 10.000 documents par an et de 3.000 à 6.000 descripteurs si la base s'étend jusqu'à 100.000 documents par an.

### *Ontologie*

L'ontologie, concept bien connu dans le domaine du web sémantique, est définie par Gruber [1993] comme « an explicit and formal spécification of a conceptualization ». La construction d'une ontologie revient donc à exprimer de manière formelle la perception que l'on a d'un domaine. Même si elle en reprend de nombreux aspects, l'ontologie déborde largement du champ d'action des *simples* terminologies.

Dans une ontologie, les concepts s'organisent en classes et disposent de propriétés. Celles-ci se rapportent à une classe ou à un type de données particulier. Cela signifie qu'il est possible de définir à peu près n'importe quel type de lien entre les différentes classes, y compris la relation hiérarchique qui est souvent nommée *is a*. Une définition logique accompagne généralement les classes et les relations, de manière à en fournir une spécification formelle, mais aussi un moyen de raisonner sur l'univers ainsi construit. Certains concepts simples ne peuvent pas être définis formellement et constituent les éléments de base de l'ontologie.

En plus des connaissances structurelles, une ontologie peut contenir des connaissances factuelles ou assertionnelles. Ces données actualisent les classes définies dans l'ontologie et sont aussi parfois appelées *instances*. Le peuplement d'instances dans le *schéma* que définit l'ontologie donne naissance à une base de connaissances. En plus des données explicites qui y ont été déposées, il est possible de déduire de la connaissance implicite à l'aide de logiciels de raisonnement. Par exemple, à partir des informations « Pierre est le fils de Jean » et « Paul est le fils de Jean », le *raisonneur* pourra déduire que Pierre et Paul sont frères.

Comme le décrit Guarino [1998], on peut distinguer plusieurs types d'ontologies. La plus générale

[32] Thésaurus du Parlement de la Communauté européenne, couvre une grande diversité de domaines, mais toujours en rapport avec le travail parlementaire : `http://europa.eu/eurovoc/`

[33] Thésaurus de l'Organisation des Nations Unies pour l'alimentation et l'agriculture, se concentre sur l'agriculture : `http://www.fao.org/aims/ag_intro.htm`

[34] Eurovoc : 6.645 pour chaque langue ; Agrovoc : 28.718 en anglais uniquement.

est appelée ontologie de haut niveau (*top-level ontology*) et décrit des concepts très généraux, indépendants d'un quelconque problème particulier. Un exemple est l'ontologie SUMO[35] (Niles et Pease [2001]). L'ontologie de domaine (*domain ontology*) et l'ontologie de tâche (*task ontology*) spécialisent toutes deux les concepts de l'ontologie de haut niveau. La première le fait pour décrire le vocabulaire d'un domaine générique (la musique[36], le domaine médical[37], etc.), alors que la seconde s'attachera à la définition du vocabulaire relié à une tâche ou à une activité générique (la pose de diagnostic, la vente, ou encore l'apprentissage assisté par ordinateur, comme chez Ikeda *et al.* [1997]). Enfin, les ontologies d'application (*application ontology*) sont les plus spécifiques. Elles définissent des concepts qui correspondent à des rôles pris dans le cadre d'une certaine activité par des entités d'un certain domaine. Il s'agit donc d'une spécialisation des types d'ontologies de tâche et d'activité.

### 1.4.2 Cas concrets d'utilisation de ressources terminologiques pour l'indexation et la recherche de documents

Dans cette section nous allons passer en revue quelques exemples qui montrent dans quels contextes sont utilisés les terminologies à des fins d'indexation et de recherche d'informations. D'une manière générale, les terminologies employées sont plutôt des thésaurus, voir des taxonomies, étant donné d'une part le coût élevé de création de structures plus complexes, et d'autre part la valeur ajoutée relative de celles-ci pour la recherche.

Le thésaurus Eurovoc[38] a été développé par la Commission européenne dans le but d'indexer les documents issus du travail parlementaire européen. Il est également utilisé par d'autres organisations telles que, en Belgique, la Chambre des députés et le Sénat. La figure 1.10, montre, pour le site du Sénat[39] les possibilités de recherche offertes suite à l'utilisation d'Eurovoc pour l'indexation. Une requête permet à la fois de restreindre l'ensemble de documents par rapport au thème à l'aide des descripteurs Eurovoc, mais aussi d'exprimer des contraintes par rapport au type de document, à sa date de publication, à ses auteurs ou encore par rapport aux mots contenus dans les titres et sous-titres.

Un deuxième exemple concerne la recherche de littérature scientifique dans le domaine des sciences de la vie sur le portail PubMed[40]. Celui-ci permet la recherche au moyen de termes issus de MeSH[41], comme illustré à la figure 1.11. La recherche se déroule en plusieurs temps : l'utilisateur introduit un mot-clé pour trouver le descripteur MeSH qu'il recherche, il le sélectionne et lance ensuite la véritable recherche de documents à partir de celui-ci. Il est possible de combiner les descripteurs à l'aide des opérateurs logiques habituels.

[35] Suggested Upper Merged Ontology (`http://www.ontologyportal.org`).
[36] Music Ontology, `http://musicontology.com`.
[37] `http://bioportal.bioontology.org`
[38] `http://eurovoc.europa.eu`
[39] `http://www.senate.be`
[40] `http://www.ncbi.nlm.nih.gov/pubmed/`
[41] Medical Subject Headings (`http://www.nlm.nih.gov/mesh/`).

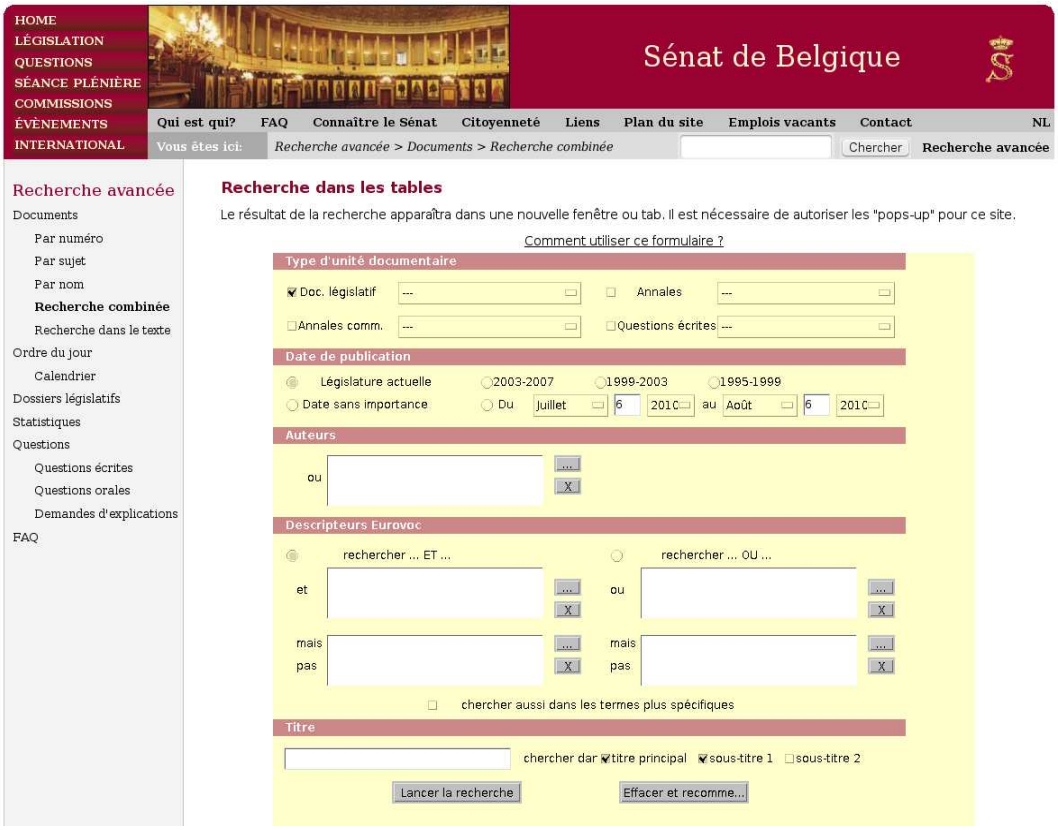


*Figure 1.10 : Le formulaire de recherche avancée du Sénat permet l'utilisation des descripteurs d'Eurovoc.*

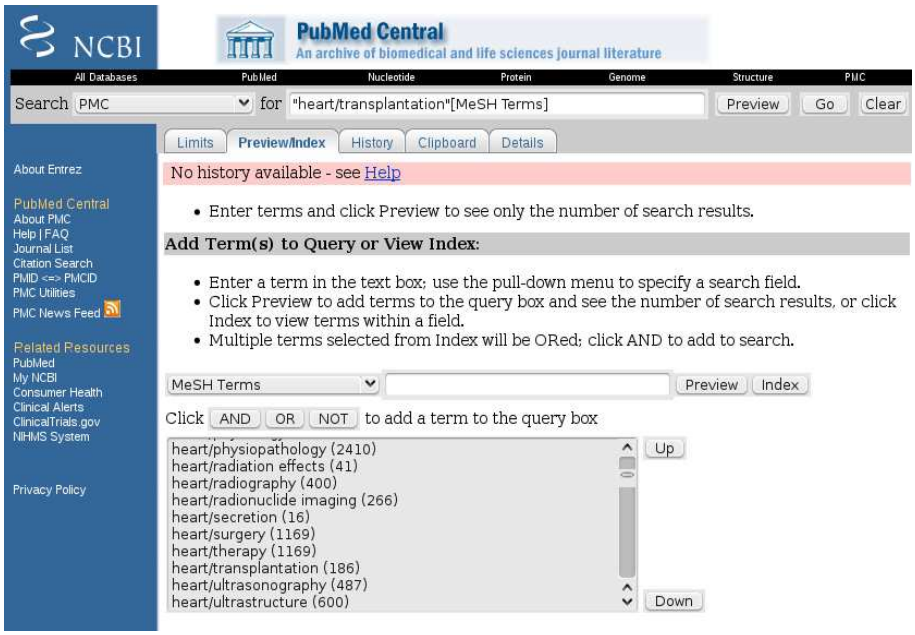


*Figure 1.11 : Utilisation de MeSH dans le formulaire de recherche avancée du Pubmed.*

### 1.4.3 Avantages, inconvénients et perspectives

L'utilisation d'un terminologie lors de l'indexation est très souvent mise en œuvre au sein d'un processus manuel. Ce sont alors des documentalistes, généralement experts d'un domaine particulier, qui analysent les textes avant de leur attribuer une ou plusieurs catégories.

Du côté de la recherche, la requête est définie par le choix d'une ou plusieurs catégories, ou par la navigation dans la structure qui les organise. Il est cependant parfois possible de faire appel à une recherche par mots clés qui détermine alors les catégories pertinentes pour cette requête. Le système d'interrogation par mots-clés pourra également être exploité pour effectuer une recherche classique sur le texte complet, tout en utilisant le système de catégories comme *filtre*. Ce dernier restreint alors la collection de documents à un sous-ensemble qui correspond à une certaine catégorie.

Le principal avantage d'une indexation au moyen d'une terminologie est qu'elle permet de contrôler l'espace d'indexation, et par là même d'y apporter un sens précis. La qualité de cette indexation, généralement manuelle, est bien connue dans les milieux documentaires (Chaumier et Dejean [2003], Da Sylva [2004]) et constitue une solution qui est implémentée dans les entreprises et autres organisations.

Cet apport sémantique de qualité est également présent lors de la recherche. En effet, chaque descripteur, ou catégorie, possède un sens et représente un concept bien précis dans le contexte thématique d'une terminologie en particulier. Une recherche sur le descripteur « carotte » n'aura pas le même sens lorsqu'il est question de légumes ou lorsque le sujet est la géologie. Dans un système utilisant une indexation par rapport à une terminologie, la recherche ne sera pas ambiguë car elle ne s'effectue pas sur le mot, mais sur un concept. Le terme est en fait implicitement désambiguïsé par les concepts plus généraux qui y sont hiérarchiquement reliés (par exemple, « légume » ou « géologie ») ou plus simplement par le fait qu'une seule interprétation existe au sein du système terminologique. Dans le cas d'une recherche par mots-clés, l'ambiguïté subsiste à tout moment.

Diverses difficultés viennent cependant tempérer ces avantages, souvent au profit des solutions d'indexation *full text* et des modes de recherche par *mots-clés* (voir la comparaison entre l'indexation humaine et l'indexation automatique, réalisée par Chaumier et Dejean [2003]). En ce qui concerne l'indexation, les coûts engendrés par l'utilisation des terminologies ne sont pas négligeables, tant en ce qui concerne la création même de la ressource que son utilisation. En effet, les documentalistes humains, experts du domaine, qui prennent souvent en charge l'attribution des descripteurs aux documents, représentent une charge financière importante. L'indexation manuelle constitue également un processus relativement lent. De plus, si cette solution apporte effectivement une bonne qualité à l'indexation, elle introduit aussi paradoxalement des problèmes de cohérence, soit au cours du temps, soit entre les différents annotateurs. Van Slype [1987] montre que la cohérence de l'indexation d'un même document par deux documentalistes se situe entre 50% et 80%. De même, Pouliquen *et al.* [2003] rapportent un *accord inter-annotateur* allant de 78% à 87%.

En ce qui concerne la recherche, des restrictions peuvent également être émises. Les interfaces d'interrogation, telles que celles qui ont été présentées à la section 1.4.2, ne sont souvent pas très satisfaisantes pour l'utilisateur. En plus de leur côté peu ergonomique, l'inconvénient principal réside, pour l'utilisateur, en sa connaissance généralement très approximative du système terminologique utilisé pour l'indexation :

> « Il est évident que dans nos dispositifs documentaires actuels, l'utilisateur final qui n'indexe pas et qui n'a pas construit le thésaurus, non professionnel de la documenta-

> tion et parfois non spécialiste du domaine, affronte en réalité une tâche beaucoup plus complexe qu'un documentaliste » (Dalbin [2007], p. 45)

L'effort nécessaire à une bonne connaissance d'une telle ressource est important et de nature à décourager de nombreux utilisateurs :

> « Mais la plupart des utilisateurs souhaitent passer outre cette phase complexe de formulation d'une requête à partir de la sélection de termes dans des vocabulaires contrôlés, préférant porter leur attention sur la fouille du lot de résultat. » (Dalbin [2007], p. 48)

Le problème provient donc du décalage qui existe, dans la maîtrise de la ressource terminologique, entre les documentalistes et les utilisateurs.

Les avantages offerts par une indexation relative à une terminologie se heurtent donc aux obstacles de coût et de lenteur de la tâche en ce qui concerne l'indexation, ainsi qu'à l'inadéquation des modes de recherche lors de l'interrogation des bases documentaires. Il existe cependant des perspectives qui rendent ce choix possible. Pour l'indexation, il s'agit principalement de mettre en place des méthodes automatiques ou semi-automatiques, selon le degré de contrôle que l'on désire conserver. Le chapitre 2 propose à cet égard un processus d'indexation semi-automatique qui utilise le thésaurus comme base d'un processus de classification dans lequel chaque descripteur, identifié par son code et accompagné par ses synonymes, constitue une classe. Pour ce qui concerne la recherche, il est possible d'exploiter l'indexation de manière beaucoup plus souple que par une navigation dans une hiérarchie de catégories. Une solution performante ne peut cependant être atteinte qu'en combinant plusieurs des techniques *sémantiques* présentées à la section 1.3.4. Par exemple, à partir d'une requête par mots-clés :

- extension de requête afin de maximiser la couverture (rappel) ;
- effectuer un regroupement de ces résultats étendus selon les catégories définies par la terminologie ;
- présenter les catégories les plus *pertinentes* afin que l'utilisateur précise sa requête (mécanisme de *relevance feedback*) ;
- le choix exprimé par l'utilisateur, permet de préciser le sens de la requête initiale, et ainsi d'atteindre un résultat sémantiquement plus proche de ses attentes (précision).

La réalisation d'un tel système n'est cependant pas effectué dans le cadre de cette thèse. Les technologies suggérées pour cette solution de recherche ont cependant prouvé leur efficacité. Leur combinaison doit permettre d'atteindre un résultat en rapport avec les exigences des utilisateurs, et encourage par conséquent le développement de systèmes (semi-)automatiques d'indexation guidés par des terminologies.