Moteurs de recherche et restitution de l'information dans les grandes entreprises : l'exemple du portail Cyberthèque de la Direction des Systèmes d'Information de la Société Générale

Alina Ivanciuc Deniau

▶ **To cite this version:**

Alina Ivanciuc Deniau. Moteurs de recherche et restitution de l'information dans les grandes entreprises : l'exemple du portail Cyberthèque de la Direction des Systèmes d'Information de la Société Générale. domain_shs.info.gest. 2003. mem_00000013

HAL Id: mem_00000013

https://memsic.ccsd.cnrs.fr/mem_00000013v1

Submitted on 17 Dec 2003

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS

INSTITUT NATIONAL DES TECHNIQUES DE LA DOCUMENTATION

MÉMOIRE PRÉSENTÉ EN VUE D'OBTENIR

LE DESS EN SCIENCES DE L'INFORMATION
ET DE LA DOCUMENTATION SPÉCIALISÉES

par **Alina IVANCIUC DENIAU**

Moteurs de recherche et restitution de l'information dans les grandes entreprises :

l'exemple du portail Cyberthèque de la Direction des Systèmes d'Information de la Société Générale

Mémoire soutenu devant un jury composé de :

Danièle DÉGEZ
Maroline LAM VAN BA

25 novembre 2003
CYCLE SUPÉRIEUR PROMOTION XXXIII

SOMMAIRE

CONVENTIONS TYPOGRAPHIQUES

[no]	=	références ; les chiffres placés entre crochets correspondent à l'ordre d'apparition dans le texte de la référence bibliographique et renvoient à la bibliographie analytique
(auteur, année, p. xy, [no])	=	références complètes ; sont présentées entre parenthèses et comprennent : le nom de l'auteur (personne physique ou personne morale), l'année de publication de l'article ou de l'ouvrage, la page (aussi souvent que possible) et la référence chiffrée (v. supra) et renvoient, d'une part, à la bibliographie analytique (via les chiffres entre crochets), d'autre part à la bibliographie alphabétique des noms d'auteurs (pour le même auteur, les articles sont classés par date, du plus récent au plus ancien ; si deux articles ou ouvrages du même auteur sont publiés la même année, ils sont notés « a », « b », etc.)
« *texte* »	=	citations : texte en italiques, en retrait, entre guillemets ; utilisé uniquement pour les citations longues (plus d'une ligne)

INTRODUCTION

« Actuellement, l'information est pléthorique et ses sources sont hétérogènes. » Les présentations de produits commercialisés par des éditeurs de logiciels sous des noms allant de « moteur de recherche » à « portail d'entreprise » en passant par « fédérateur de requêtes » commencent souvent par ce truisme. On ne peut nier la multiplication des sources d'information, ni la diversité des formats de fichiers plus ou moins (in)compatibles entre eux, toujours est-il que la nécessité d'avoir la bonne information au bon moment reste une des données vitales dans le quotidien des entreprises.

Moteurs de recherche et autres produits commercialisés sous des bannières relevant plus ou moins de la stratégie marketing sont aujourd'hui des briques logicielles de plus en plus répandues dans les entreprises, surtout en ce qui concerne les grands comptes. Leur intégration dans des architectures informatiques diverses et variées (client-serveur, intranet, serveur d'application) ne se fait pas toujours sans douleur. Selon toute vraisemblance, les intégrateurs ont de beaux jours devant eux, car l'harmonisation des systèmes, des architectures et des modes de fonctionnement de chaque service particulier ne peut se faire sans une volonté forte des dirigeants ; les coûts d'une telle opération sont difficiles à supporter par une grande entreprise dans un contexte économique plutôt défavorable. D'autre part, les applications déjà présentes dans une entreprise ou dans un service particulier ont quelquefois une identité graphique forte, qu'il n'est pas souhaitable d'abandonner au profit d'une interface quelquefois dépouillée ou, bien au contraire, d'une couleur trop soutenue ou offrant des fonctions que le cahier des charges n'a pas retenues.

En attendant, l'information circule et elle doit circuler sous peine de sclérose du système. Retrouver l'information stockée dans l'entreprise est une composante importante de la recherche d'information au quotidien, ne serait-ce que pour retrouver, par exemple, la nouvelle procédure d'attribution des primes, ou la nouvelle norme concernant telle ou telle application particulière qu'un groupe de travail souhaite intégrer dans un nouveau projet. Les solutions logicielles existent, le marché des moteurs de recherche est en plein mouvement (rachats, émergence de produits nouveaux) et les entreprises investissent de plus en plus, depuis quelques années, dans l'achat (sans oublier la maintenance) de moteurs de recherche.

Mais pour les appréhender, les choisir et les utiliser à bon escient et, surtout, pour éviter les déceptions, une certaine compréhension de leur fonctionnement semble nécessaire, voire indispensable.

❑ Moteurs de recherche et recherche d'information

Les moteurs de recherche sont des programmes informatiques complexes issus de la recherche en traitement automatique des langues (TALN). Leur développement et leur mise au point font appel à des disciplines multiples et à des techniques sophistiquées, impliquant la linguistique, l'informatique, la logique, les mathématiques, la statistique et les sciences cognitives.

Après avoir replacé ces outils dans le contexte de la recherche d'information et des langages documentaires (notamment le thésaurus), compte tenu des mutations de ces dernières années (taxonomies, ontologies, Topic Maps), la première partie se propose de décrire le fonctionnement des moteurs de recherche. La définition du traitement automatique des langues naturelles (TALN), en soulignant les apports de chaque discipline avec un éclairage particulier sur la linguistique, et une typologie des produits présents dans les grandes entreprises seront suivies de la description proprement dite du fonctionnement des moteurs de recherche. Les opérations effectuées par les moteurs de recherche linguistiques pour traiter la masse d'information afin de répondre à la requête de l'utilisateur prendront pour exemple un produit particulier : K2 Enterprise de la société Verity.

Cette présentation a pour objectif de situer dans un contexte précis les attentes qu'un responsable souhaitant acquérir un tel outil pourrait trouver déçues une fois le système mis en marche. Afin d'éviter cet écueil, il convient peut-être d'avoir des attentes raisonnables, fondées sur une relative compréhension du fonctionnement des outils. Il serait possible ainsi de se faire une idée plus précise de la concordance entre les besoins du service (et des utilisateurs finaux) et les produits commercialisés, du temps nécessaire au paramétrage du moteur ainsi que des paramètres à prendre en compte en fonction des contraintes spécifiques au service ou au secteur dans lequel il exerce.

La compréhension de l'outil, conjuguée à la connaissance du fonds et du public, peut contribuer à améliorer la gestion d'un projet de mise en place (en facilitant la formalisation du cahier des charges), la prise de décision concernant les paramétrages possibles et la gestion au quotidien d'un moteur de recherche dans un service documentaire, veille (ou autre).

❑ Audit : Verity K2 à la Société Générale

La seconde partie retrace l'audit effectué afin de mettre en place des améliorations dans un portail d'entreprise dont la recherche est gérée par le moteur de recherche Verity K2. La Cyberthèque de la Direction des Systèmes d'Information de la branche Banque de Détail de la Société Générale est un portail de veille technologique et concurrentielle.

Les exemples figurant dans la première et la deuxième partie ont pour origine l'application Verity dans le portail Cyberthèque.

PREMIÈRE PARTIE :

1 MOTEURS DE RECHERCHE

1.1 LES MOTEURS DE RECHERCHE ET LA RECHERCHE D'INFORMATION

La problématique de la recherche d'information sur les réseaux est, depuis quelques années, au centre des préoccupations des professionnels de l'information (Maniez, 1999, [1] ; Le Moal, 2002, [2]).

Le Web sémantique, théorisé par Tim Berners-Lee, est une tentative de structuration de l'information disponible sur le grand réseau mondial, Internet.

Quant aux réseaux internes des entreprises, les intranets, plus ou moins structurés, plus ou moins reliés entre eux, ils drainent une grande quantité d'information stockée dans l'entreprise. Est-elle pour autant plus facile à retrouver pour les employés ? Rien n'est moins sûr.

Une étude du cabinet d'analystes IDC publiée en juillet 2001, intitulée « The High Cost of Not Finding Information » (IDC, 2001, [3]), fait le calcul de l'argent perdu par les entreprises à cause de l'information non-trouvée.

Le constat est à prendre en compte, mais, d'autre part, il n'est pas nouveau. En effet, une loi, connue depuis longtemps en sciences de l'information (Lefèvre[1], 2000, p. 53, [4]), s'exprime ainsi :

> *« Tout travail de classement et de référencement non réalisé en amont, au moment de la réception d'information, se traduit ensuite, en aval, au moment de la recherche, par une dépense d'énergie, un temps et un coût supérieurs de plusieurs ordres de grandeur. »*

Les moteurs de recherche résoudront-ils le problème ? Leur succès grandissant et leur médiatisation récente semblent appeler à répondre par l'affirmative.

Mais la recherche en texte intégral (full-text) s'avère vite insatisfaisante pour l'utilisateur lambda, car il faut se former à l'utilisation des opérateurs de requête pour obtenir des résultats pertinents.

On en revient donc au constat antérieur : pour retrouver l'information, il faut la ranger dès son acquisition (Lefèvre, 2000, p. 53, [4]) :

> *« La présence d'outils d'indexation et de recherche sur le contenu pourrait faire croire qu'il est possible de se passer d'une organisation préalable de l'information, et que les moteurs de recherche pallieront à ce désordre. L'expérience prouve que c'est faux. »*

Proposer des modules destinés à organiser l'information est la tendance actuelle des outils logiciels proposés par les éditeurs de moteurs de recherche. C'est aussi la voie montrée par les ontologies, briques essentielles du Web sémantique.

[1] Cite une étude de Christian Fluhr datant de 1992.

Les modules de catégorisation (automatique ou supervisée) et/ou de classification automatique inclus dans les offres actuelles des éditeurs sont des outils appropriés pour la construction de terminologies (référentiels d'entreprise, taxonomies, thésaurus, etc.) adaptées aux besoins des entreprises.

Mais les besoins et les pratiques informationnelles des entreprises ont beaucoup évolué ces dernières années. Un aperçu historique de la gestion documentaire s'impose.

1.1.1 GESTION DOCUMENTAIRE

Pour avoir un aperçu historique de la gestion documentaire dans les entreprises au cours du XXe siècle, selon Jacques Maniez (Maniez, 2002, p. 160, [5]), on peut établir trois étapes successives :

- **Période « tout papier »**

La première période, qui dure jusque dans les années 1960, serait celle de la gestion entièrement papier, utilisant les tiroirs de fiches afin de stocker les notices catalographiques qui renvoient aux documents papier disponibles sur place. La recherche s'effectue en feuilletant les fiches.

- **Centre de documentation informatisé**

La seconde période, de 1960 aux années 1990, est celle de la documentation imprimée informatisée. L'essor de la micro-informatique (à partir de 1980) est suivi de l'essor des logiciels documentaires fonctionnant en réseau client-serveur. La plupart possèdent des moteurs de recherche internes qui permettent de trouver et de trier les notices. La recherche s'effectue selon des critères plus ou moins complexes, utilisant les opérateurs booléens (au minimum) et les opérateurs de proximité. C'est la période de l'essor des langages documentaires (voir partie 1.1.2 LANGAGES DOCUMENTAIRES, page 9).

- **Documentation numérique**

Deux cas de figure sont possibles : les documents sont scannés ou les documents d'origine sont en format numérique. Mais la caractéristique fondamentale tient à l'architecture des applications qui passent du mode client-serveur à l'architecture intranet, voire à une architecture distribuée, où l'accès aux sources se fait indifféremment de l'endroit (unique ou multiple) dans lequel les documents électroniques sont stockés. À cela s'ajoute l'information disponible sur Internet. La recherche devient un problème à gérer au jour le jour, car la facilité d'accès à une masse d'information de plus en plus importante demande de plus en plus de temps pour la rechercher et la trier.

1.1.2 LANGAGES DOCUMENTAIRES

Les langages documentaires sont des constructions intellectuelles qui visent à formaliser la connaissance d'un domaine particulier à l'aide des termes de spécialité rencontrés et des relations entre ces notions (voir à ce sujet : AFNOR, 1987, [6] ; Le Coadic, 1997, [7] ; Dégez, 2001, [8] ; EBSI, 2002, [9]). Leur but est d'offrir un maximum de concordance entre la description du contenu d'un document X et le terme qu'un utilisateur lambda pourrait utiliser lors d'une recherche visant à obtenir des renseignements au sujet d'une notion ou d'un concept particulier appartenant au domaine dont il est question dans ce document. Petite précision : le terme ne se trouve pas forcément dans le titre, ni dans le texte (résumé ou document primaire scanné). La présence du champ traditionnellement appelé « Indexation », « Mots-clé » ou « Descripteurs » dans les logiciels documentaires permet donc d'élargir la recherche à des textes qui y échapperaient en son absence. Le même cas de figure se présente dans les logiciels des bibliothèques, qui peuvent gérer, en plus de la liste des mots-clés autorisés (listes d'autorité, RAMEAU), une classification à vocation universelle (Dewey ou CDU) qui détermine en même temps l'« adresse » physique du document dans la bibliothèque, surtout si celle-ci est adepte du libre service et possède peu de fonds stockés en magasin.

Laissons de côté l'utilisation du langage naturel pour décrire le fonds documentaire, une place trop importante peut être accordée à la fantaisie dans la saisie des mots-clés. Aucun contrôle de la saisie n'est possible en l'absence de toute liste normalisée des termes autorisés.

Les langages contrôlés ou langages documentaires, selon la terminologie recommandée par l'AFNOR (AFNOR, 1987, [6]), sont de type hiérarchisé (classification, nomenclature, plan de classement, taxonomie) ou de type combinatoire (lexique, liste d'autorité, thésaurus). Ce sont des langages artificiels, au même titre que les langages informatiques, dans le sens où il ne s'agit pas de langues (Natural Language en anglais), mais de langages créés et normalisés dans un but précis : celui d'éliminer les ambiguïtés et la redondance spécifiques au langage naturel lors de l'indexation des documents (voir partie 1.1.3 INDEXATION MANUELLE ET INDEXATION AUTOMATIQUE, page 11).

Quelques définitions à caractère officiel sont rappelées ici :

⇒ Langage documentaire

Langage artificiel constitué de représentations de notions et de relations entre ces notions et destiné, dans un système documentaire, à formaliser les données contenues dans les documents et dans les demandes des utilisateurs (AFNOR, 1987, p. 72, [6]).

⇒ Langage contrôlé

Langage documentaire comprenant des termes d'indexation et leurs règles d'utilisation (Voir langage documentaire et voir langage artificiel) (AFNOR, 1987, p. 72, [6]).

⇒ Langage artificiel
Langage construit ou contrôlé à l'aide d'un ensemble de règles (AFNOR, 1987, p. 72, [6]).

Les langages documentaires peuvent être hiérarchisés ou combinatoires :

❑ Langages hiérarchisés

⇒ Classification
Langage documentaire fondé sur la représentation structurée d'un ou plusieurs domaines de la connaissance en classes et dans lequel les notions et leurs relations sont représentées par les indices d'une notation (AFNOR, 1987, p. 39, [6]).

⇒ Nomenclature
Classification méthodique de l'ensemble des termes d'un domaine spécialisé (AFNOR, 1987, p. 84, [6]).

⇒ Plan de classement
Document qui présente une classification de manière ordonnée, en faisant apparaître la signification donnée à chaque indice, et le cas échéant les relations entre les classes. Le plan de classement peut également inclure des recommandations ou des consignes quant à l'utilisation de la classification (Dégez, 2001, p. 33, [8]).

⇒ Taxonomie
Classification des formes vivantes (Dégez, 2001, p. 41, [8]).

❑ Langages combinatoires

⇒ Lexique
Liste de mots d'une ou plusieurs langues dans un domaine donné (AFNOR, 1987, p. 73, [6]).

⇒ Liste d'autorité
Liste des vedettes ou termes qui doivent être obligatoirement et nécessairement utilisés dans le catalogage ou l'indexation (AFNOR, 1987, p. 74, [6]).

⇒ Thésaurus
Langage documentaire fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et les relations entre notions par des signes conventionnels (AFNOR, 1987, p. 112, [6]).

En conclusion, dans la mesure où il s'agit d'adapter la pratique des langages documentaires à des logiciels comme les moteurs de recherche, dont le fonctionnement ne permet pas toujours de gérer des relations de type « voir aussi » (terme associé ou Related Term en anglais) le terme préférentiel employé pour langage documentaire sera langage contrôlé.

1.1.3 INDEXATION MANUELLE ET INDEXATION AUTOMATIQUE

Les index et l'indexation sont définis différemment selon le domaine dont on parle : édition, informatique, documentation :

⇒ édition

Dans le domaine de l'édition, l'index d'un livre, situé généralement à la fin, est une liste des termes choisis par l'auteur, considérés comme significatifs, accompagnés des numéros des pages où ils apparaissent.

⇒ informatique

En informatique (Lefèvre, 2000, p. 105, [4]), un fichier index sert surtout de pointeur :

> *« Dans une base de données on appelle fichier index, un fichier qui comprend un élément pour chaque enregistrement logique de la base de données, et dont chaque élément est constitué de deux zones : la clé d'enregistrement logique, et un pointeur, qui indique son adresse dans la base de données. »*

⇒ documentation

Un index, dans le domaine de la documentation, est défini (AFNOR, 1987, p. 67, [6]) comme :

> *« Liste ordonnée de noms de personnes, de lieux et de matières figurant dans un document assortis d'une référence permettant de les retrouver. »*

L'indexation, dans le domaine de la documentation, est définie (AFNOR, 1987, p. 67, [6]) ainsi :

> *« Processus destiné à représenter par des éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou d'une question. On désigne également ainsi le résultat de cette opération. »*

Dans le cas de l'indexation manuelle, l'accent est mis sur la correspondance entre la « formalisation du domaine de connaissances (thésaurus) » et la représentation du contenu du document (Jolion, 2000, p. 139, [10]).

L'indexation automatique « utilise diverses méthodes d'analyse appliquées au texte intégral pour représenter le contenu du document » (Jolion, 2000, p. 139, [10]).

Qu'elle soit manuelle ou automatique, sur un grand réseau comme Internet (Metzger, 2001, [11]) ou sur un corpus restreint, l'indexation a pour but de permettre de retrouver un document en fonction de son contenu informationnel.

Même si Muriel Amar (Amar, 2000, pp. 26-28, [12]) conteste l'approche « instrumentale » de l'indexation dans le modèle Information Retrieval[2], ce point de vue est probablement le plus approprié en ce qui concerne le fonctionnement des intranets des entreprises. En effet, il s'agit, le plus souvent, de retrouver de l'information connue ou supposée telle, dont on ne connaît pas l'emplacement exact, ni la teneur exacte (ex. nouvelle façon d'attribution des congés, mise en ligne par le service des ressources humaines), mais dont on suppose l'existence. L'indexation effectuée par les moteurs de recherche est destinée à faciliter la restitution[3] de l'information présente dans le système.

En conclusion, pour toutes ces raisons, avant de décrire le fonctionnement des moteurs de recherche, il est nécessaire de faire le point sur les systèmes créés pour gérer l'information, à travers une typologie des systèmes d'organisation des connaissances, notamment ceux dont on parle le plus actuellement : taxonomies, thésaurus, Topic Maps et ontologies.

La façon de les intégrer dans des moteurs de recherche présents sur le marché sera discutée ensuite, en prenant comme exemple le cas de Verity K2 et de son module de gestion des « concepts » et de la taxonomie, Intelligent Classifier.

[2] Dans l'analyse des pratiques de recherche d'information, le modèle « Information Retrieval » consiste à considérer que la recherche porte sur de l'information connue à retrouver. S'oppose au modèle « Search », d'influence cognitiviste, où on suppose chercher de l'information sans savoir si elle existe, ni si elle est disponible (comme sur le web) et, donc, l'accent est mis sur la nécessité de construire une stratégie de recherche.

[3] C'est dans ce sens que le terme « restitution » est utilisé dans le titre : il s'agit de retrouver l'information stockée dans l'entreprise, d'où le syntagme « restitution de l'information ». La notion qui n'est pas définie dans le champ des sciences de l'information, mais « *retrouvage » n'existe pas en français.

1.2 LES SYSTÈMES D'ORGANISATION DES CONNAISSANCES

Sous le nom de « systèmes d'organisation des connaissances » seront regroupés tous les langages contrôlés qui peuvent être traités par la machine dans la perspective de leur usage sur le Web sémantique. Le terme est la traduction de l'anglais « Knowledge Organization Systems » (KOS)[4] (Soergel, 2003, [13] ; Smith, 2003, [14]). Il désigne autant les langages documentaires classiques de type hiérarchique (classifications, nomenclatures, plans de classement, taxonomies) ou combinatoire (lexiques, listes d'autorité, listes de vedettes-matières, thésaurus), que les ontologies et Topic Maps, formes nouvelles de représentation des connaissances dont la mise en oeuvre informatique est fortement structurée.

Les langages documentaires servent au contrôle de l'indexation afin d'assurer son homogénéité et sa cohérence lors de l'indexation manuelle. À l'autre bout de la chaîne, en recherche, leur adéquation avec les besoins des usagers garantit la restitution[5] (ou la récupération) des documents pertinents.

Les Topic Maps et les ontologies, nouveaux outils de contrôle du langage d'indexation et de représentation des connaissances, revendiquent une caractéristique fondamentale en termes de gestion électronique des documents : la séparation du document électronique et de la représentation conceptuelle. Les façons de pointer vers le document électronique sont décrites dans des normes.

Les avantages de cette séparation sont multiples : décrits comme réutilisables, adaptables, portables, évolutifs, les nouveaux systèmes d'organisation des connaissances ont un grand potentiel de développement.

Les taxonomies utilisées pour la catégorisation / classification automatique des documents par certains moteurs de recherche (comme Verity K2) fonctionnent de la même manière : la taxonomie est indépendante de la collection de documents sur laquelle s'effectue la recherche, mais, utilisant des formats propriétaires, sa portabilité reste limitée.

Utilisés quelquefois comme synonymes, taxonomies et ontologies (Adams, 2002, p. 22, [15] ; Ramos, 2002, p. 2, [16]) présentent des différences fonctionnelles qui seront discutées.

[4] Le terme est utilisé par Dagobert Soergel, Gail Hodge, Marcia Lei Zeng dans le cadre du NKOS Workshop (Network Knowledge Organization Systems). La réflexion de ce groupe de travail porte sur la transformation des outils classiques de contrôle du vocabulaire, tels les plans de classement et thésaurus pour les faire évoluer vers des outils tels les ontologies et les Topic Maps, qui présentent au moins deux avantages : conceptualisation plus fine des relations gérées et formalisation pour le traitement par ordinateur. En mai 2003, le groupe de travail organisait son 6e atelier, NKOS Workshop, en marge de la conférence ACM-IEEE Joint Conference on Digital Libraries (JCDL).

[5] Dans le sens de retrouver une information présente dans une base de connaissances (système documentaire, base de données, intranet). C'est dans ce sens que le mot est utilisé dans le titre du mémoire.

❑ Réseaux sémantiques

Taxonomies, thésaurus, Topic Maps, ontologies : toutes ces structures s'appuient sur des réseaux sémantiques. Un réseau sémantique est un ensemble de mots représentant des concepts ou des objets liés par des relations sémantiques.

Représentations élaborées à l'origine dans le domaine de la psychologie expérimentale, par R. Quillian, en 1968, afin de rendre compte de la façon dont les humains catégorisent et mémorisent les concepts, les réseaux sémantiques ont été utilisés dans le domaine de l'intelligence artificielle (discipline qui tente de modéliser le raisonnement et la mémoire humaine) afin de formaliser la connaissance dans les systèmes experts et les systèmes à base de connaissances (Knowledge Based Systemes ou KBS).

Dans la terminologie de l'intelligence artificielle, les réseaux sémantiques sont formellement des graphes, constitués de « noeuds » représentant les concepts, reliés par des « arcs » étiquetés et orientés qui expriment les relations sémantiques entre les noeuds (Lefèvre, 2000, p. 136, [4]).

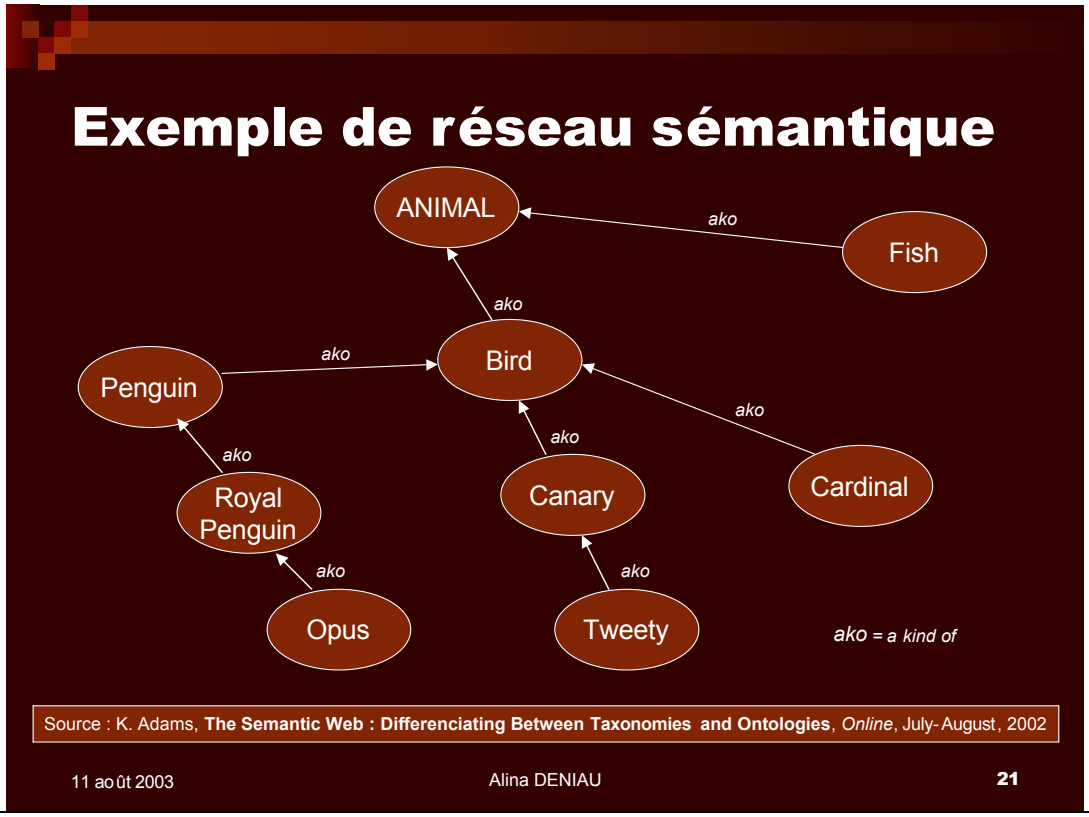
Les graphes conceptuels, théorisés en 1984 par John Sowa en tant que modèles de représentation des connaissances, peuvent être rapidement définis (Metzger, 2001, p. 201, [11]) ainsi :

> *« Le modèle des graphes conceptuels est un modèle de représentation de connaissances de type « réseaux sémantiques » permettant une représentation sous forme graphique. »*

Les réseaux sémantiques servent à représenter divers types de connaissances dans différents domaines par le biais de relations.

Un réseau sémantique (voir Figure 1 - Exemple de réseau sémantique simple (taxonomie), page 15) met en oeuvre des relations de type :

- ⇒ hiérarchique : « sorte-de » (hyponymie-hyperonymie), relation d'inclusion de classes
- ⇒ relation tout-partie : « partie-de » (holonymie-méronymie)
- ⇒ synonymie (et quasi-synonymie non-transitive)
- ⇒ antonymie : « contraire-à »
- ⇒ « sert-à »
- ⇒ « conséquence-de »
- ⇒ « fait-en »
- ⇒ ...



Source : Katherine Adams, The semantic Web : Differentiating Between Taxonomies and Ontologies, Online, July-August, 2002, p. 22 (Adams, 2002, [15])

Figure 1 - Exemple de réseau sémantique simple (taxonomie)

1.2.1 TYPOLOGIE DES SYSTÈMES D'ORGANISATION DES CONNAISSANCES

1.2.1.1 TAXONOMIES

Le terme « taxonomie » est de plus en plus utilisé pour désigner les classifications et plans de classement mis en place pour organiser le contenu et faciliter la navigation et la recherche sur les sites Internet et intranet.

Le Petit Robert et Le Petit Larousse conseillent l'usage du terme « taxinomie » pour l'acception usuelle, apparentée à la biologie et à la classification en général. Pour l'organisation de l'information sous format électronique le terme retenu ici sera donc taxonomie.

Un exemple récent de structuration de site à l'aide d'une taxonomie construite en fonction des besoins, servant d'aide à la navigation, est le site e-business www.egreetings.com (Farnum, 2002, [17]). La construction d'une taxonomie des produits dérivés du personnage de Snoopy dessiné par C. Schultz a été réalisée (Bertolucci, 2003, [18]) pour la société qui en détient les droits. D'autres exemples semblent montrer que le monde de l'entreprise découvre les vertus de la classification décimale de Dewey (Ainsbury, 2002, [19]). Il existe même un projet de structuration de la navigation sur le web à l'aide de la classification de Dewey et du thésaurus IEEE du web (Saeed, 2002, [20]).

Le terme vient du monde de la biologie (classification hiérarchique des espèces vivantes par Linné, à la fin du XVIII^{e} siècle) mais désigne actuellement surtout la démarche classificatoire adoptée : allant du général vers le particulier. Les grandes classes constituées se subdivisent en classes de plus en plus petites, selon des critères de plus en plus pointus. C'est aussi la démarche des grandes classifications décimales utilisées dans les bibliothèques : la Classification décimale de Dewey (créée par Melvill Dewey en 1873 et publiée en 1876) et la Classification décimale universelle (CDU), adaptation de la précédente (1910).

Les plans de classement et les nomenclatures (langages hiérarchiques) sont conçus selon la même approche (pour les définitions, voir partie 1.1.2 LANGAGES DOCUMENTAIRES, page 9).

Les taxonomies permettent la gestion de **relations hiérarchiques** relativement floues, ce qui implique une grande liberté dans la démarche de construction (ce qui est un avantage), mais le résultat reste tributaire aux diverses interprétations possibles lors de leur utilisation en recherche, ce qui est un désavantage en comparaison avec des systèmes d'organisation des connaissances plus finement structurés.

Les relations entre les termes sont des relations hiérarchiques, de type « est une sorte de » (« is a kind of », générique), « est un » (« is a », instanciation), et « est une partie de » (« is a part of », partitive), mais aucune différence n'est faite entre les types de relations.

L'expression graphique d'une taxonomie est un arbre (une arborescence) ou un système de répertoires, allant du général vers le particulier.

Dans le monde de l'entreprise, les taxonomies sont définies comme des structures hiérarchiques, abstraites et ordonnées systématiquement destinées à la classification des concepts ou des objets (Ramos, 2003, p. 1, [21]) :

> *« Taxonomy represent hierarchically ordered, systematic and abstract structure for the classification of concepts or things. »*

Intégrées à des systèmes plus complexes de type portail, Web Content Management ou Data Mining, les « corporate taxonomies » (Gilchrist, 2001, [22]) servent à organiser le contenu et à aider les différents utilisateurs (employés, partenaires, clients) à naviguer dans un contexte spécifique au secteur d'activité.

Le but principal est de retrouver plus facilement l'information nécessaire à l'exercice de la profession de chaque utilisateur. Plus concrètement, les objectifs de la mise en place d'une taxonomie dans une entreprise peuvent être :

- ⇒ recherche d'information et désambiguïsation des résultats
- ⇒ organisation du contenu et aide à la navigation
- ⇒ identifier les compétences et intérêts de chaque utilisateur (personnalisation ou « réseaux sociaux » pour Verity K2)
- ⇒ organisation des projets, processus, etc. par type
- ⇒ préparation du contenu pour la visualisation

Les éditeurs de solutions de Content Management et les éditeurs de moteurs de recherche intègrent dans leurs offres des modules de construction de la taxonomie (comme Intelligent Classifier pour Verity K2). D'autres moteurs de recherche génèrent eux-mêmes une taxonomie des notions rencontrées en s'appuyant sur les documents (Autonomy, Inxight).

Quant à la question qui commence à se poser s'il faut acheter une taxonomie ou la construire (Knox, 2003, [23]), il semble que les réponses restent mitigées. L'avantage est de ne pas partir de rien, mais toute taxonomie disponible sur le marché, via les agrégateurs de contenu (LexisNexis, Factiva) et les éditeurs de logiciels (comme Documentum, ou Convera), ou du domaine public (dmoz.org, taxonomywarehouse.com) demande des aménagements pour correspondre à l'usage spécifique de chaque service ou entreprise.

Des entreprises comme British Broadcasting Corporation (BBC), Glaxo Wellcome, Microsoft ou Unilever ont mis en place des taxonomies pour organiser le contenu de leurs portails d'information. Les solutions adoptées sont discutées par Alan Gilchrist dans le cadre d'une étude plus large portant sur 22 entreprises, suite à une enquête effectuée en 2000 (Gilchrist, 2001, [22]).

1.2.1.2 THÉSAURUS

Le thésaurus apparaît dans les années 1960, à la même époque que l'informatique. Le terme a été utilisé pour la première fois en 1957, par Hélène Brownson et le premier thésaurus est celui de la société Du Pont de Nemours (1959), suivi par le « Thesaurus of ASTIA Descriptors » (États-Unis, Département de la Défense) édité pour la première fois en 1960 et comprenant 8.000 termes. En 1964, EJC (Engeneering Joint Council) édite son thésaurus, qui sera fusionné avec celui de l'ASTIA, et formeront ainsi le thésaurus TEST (plus de 25.000 termes) (Chaumier, 2003b, p. 35, [24]).

Les thésaurus sont décrits par des normes internationales ISO (International Organization for Standardization — l'Organisation internationale de normalisation) adaptées au niveau national par certains pays. En France, l'organisme attitré est l'Association Française pour la normalisation (AFNOR), aux États-Unis, le National Information Standard Institute (NISO), organisme accrédité par American National Standard Institute (ANSI), en Grande Bretagne, le British Standards Institute (BSI).

Les normes décrivent les relations et la notation, ainsi que la présentation des thésaurus à l'écran et sous format papier.

Les normes qui régissent la construction des thésaurus sont :

Normes ISO :
- ISO 2788 : 1974 (révisée en 1986) — Principes directeurs pour l'établissement de thésaurus monolingues
- ISO 5964 : 1985 — Principes directeurs pour l'établissement et le développement des thésaurus multilingues

Normes AFNOR :
- NF Z 47-100 (1981) — Règles d'établissement des thésaurus monolingues
- NF Z 47-101 (1990) — Principes directeurs pour l'établissement des thésaurus multilingues
- NF Z 47-103 (1980) — Thésaurus monolingues et multilingues — Symbolisation des notations

Normes BS (Grande Bretagne) :
- BS 5723:1987 — Guidelines for the establishment and development of monolingual thesauri
- BS 6723:1985 — Guidelines for the establishment and development of multilingual thesauri

Une nouvelle norme britannique est en cours de réalisation :
- BS 8723 — Structured vocabularies for information retrieval — Guide

Cette nouvelle norme sera destinée aux vocabulaires structurés : « standard for structured vocabularies » (Dextre Clarke, 2003, [25]). L'accent est mis sur le traitement informatique des thésaurus et des vocabulaires autres que les thésaurus et sur l'interopérabilité entre les vocabulaires et entre les applications qui les

prennent en charge : moteurs de recherche, systèmes de Content Management, web publishing etc.

Norme ANSI-NISO (États-Unis) :

- Z39.19-2003 (1974, révisée : 1980, 1988, 1993, 1998) — Guidelines for the Construction, Format and Management of Monolingual Thesauri

La norme américaine vient d'être révisée à nouveau et approuvée le 28 août 2003. Un débat portant sur les mêmes sujets que ceux concernant la nouvelle norme britannique en cours de développement avait animé la commission de révision (Warner, 2003, [26]).

La définition du thésaurus selon la norme AFNOR NF Z 47-100 (1981) est la suivante :

> *« Un thésaurus est une liste d'autorité organisée de descripteurs et de non-descripteurs obéissant à des règles terminologiques propres et reliés entre eux par des relations sémantiques (hiérarchiques, associatives ou d'équivalence). Cette liste sert à traduire en langage artificiel dépourvu d'ambiguïté des notions exprimées en langage naturel. »*

Les thésaurus gèrent des réseaux de concepts, exprimés par un terme préférentiel (appelé « descripteur ») vers lequel renvoient les autres termes qui expriment la même notion (appelés « non-descripteurs »). Les relations entre les concepts ne sont que partiellement étiquetées (hiérarchique, synonymie, associative), mais lors de la phase de conception d'un thésaurus, tous les types de relations prises en compte sont validées à l'aide de règles de contrôle par l'équipe qui le construit.

Dans les thésaurus, trois types de relations sémantiques sont nommées explicitement :

- **Relations hiérarchiques**

La relation hiérarchique est une relation ambivalente qui lie un terme général (TG) à un terme spécifique (TS). Elle peut être de type générique (notée TGG - TGS en français) ou partitive (TGP - TSP), mais la distinction est facultative. Un autre type de relation hiérarchique est l'instanciation (objet, souvent nom propre, représentant une instance d'une classe, exprimée par un nom commun) dont la notation (ex. : BTI - NTI dans la norme américaine) est aussi facultative.

Le schéma des relations hiérarchiques, spécifié par les normes, est obligatoirement de type : grand-père — père — fils.

- **Relations d'équivalence (synonymie)**

La relation d'équivalence s'établit entre le terme préférentiel choisi pour décrire le concept, appelé « descripteur », et le (ou les) terme(s) qui ont la même signification, appelés « non-descripteurs ». Le descripteur sera utilisé pour l'indexation des documents, alors que le (ou les) non-descripteur(s) sera (seront) systématiquement renvoyé(s) vers le descripteur (« voir », noté EM - EP).

❑ Relations associatives

La relation d'association (« voir aussi »), est plus floue et peut refléter une manière subjective d'envisager le problème. Elle permet aussi d'éviter la polyhiérarchie (subordination d'un terme spécifique à plusieurs termes génériques), déconseillée dans la plupart des cas, inévitable dans d'autres.

Les types de rapprochements pris en charge par la relation associative dans un thésaurus sont de type : cause - son effet, tout - partie, action - son agent, action - son produit, action - son objet, action - lieu de l'action, science - son objet, objet - sa propriété, objet - son application, matériau - produit (Hudon, 1994, [27]).

Un autre aspect important des thésaurus sont les facettes, qui permettent d'organiser les concepts exprimés par des termes non seulement selon les thèmes (qui dépendent du domaine), mais aussi par facettes comme : processus, outil, propriété, phénomène, équipement, être vivant, matériau (Hudon, 1994, p. 84, [27]), qui sont applicables à tous les domaines.

Outil destiné au départ à un usage manuel, le thésaurus est bien géré par certains logiciels de gestion de bases de données documentaires. Mais la formalisation informatique des relations gérées par un thésaurus n'est pas définie par les normes en vigueur et l'interopérabilité des différents logiciels n'est pas forcément garantie.

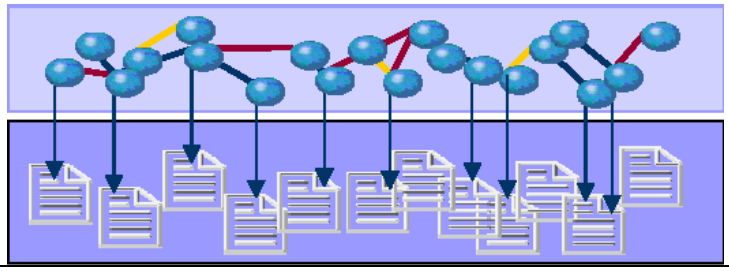
Mais, en échange, la démarche de construction des thésaurus est formalisée et appliquée depuis longtemps avec succès dans différents domaines de la vie des entreprises. Concernant l'aspect construction de thésaurus, voir la partie 1.2.3 DÉMARCHES DE CONSTRUCTION, page 28.

Certains moteurs de recherche, comme Verity K2 Enterprise (version 4.0.1), permettent d'établir des relations hiérarchiques et d'équivalence (via une interface de gestion des concepts et de la taxonomie), mais ces relations ne correspondent pas parfaitement aux spécifications des normes de création et de gestion de thésaurus, car il n'est pas possible d'avoir un terme préférentiel.

1.2.1.3 TOPIC MAPS

Nouveau type de structures liées au Web sémantique, les Topic Maps fonctionnent comme des systèmes sophistiqués de taxonomies assumant multiples points de vue et pointant (voir ci-dessous, Figure 2 - Séparation des documents et des Topic Maps) vers les documents (Adams, 2002, p. 23, [15]) :

> « *Topic maps function as a super-sophisticated systems of taxonomies, defining a group of subject and then providing hypertext links to text about these topics. Topic Maps lay out a structured vocabulary and then point to documents about these topics.* »



Source : Mondeca, Making sense of content, 2002, p. 3 (Mondeca, 2002, [28])

Figure 2 - Séparation des documents et des Topic Maps

La première version de la norme internationale concernant les Topic Maps a été adoptée par l'Organisation internationale de normalisation (ISO — International Organization for Standardization) et la Commission électrotechnique internationale (IEC — International Electrotechnical Commission) en février 2000. La seconde édition date de mai 2002 et la version en cours a été adoptée en mai 2003 :

- ISO/IEC 13250 : 2003 — Information technology — SGML applications — Topic Maps

La traduction française de « Topic Maps », présente sur la page de garde de la seconde version (mai 2002) de cette norme disponible pour l'instant uniquement en anglais, est « cartes topiques ».

À la différence des normes concernant les thésaurus, qui décrivent les relations, la démarche de construction et la forme du thésaurus à l'affichage et à l'impression, dans le but de l'utiliser en indexation manuelle, la norme ISO/IEC 13250 est destinée au traitement informatique de corpus importants et fournit des spécifications quant à l'implémentation des applications. C'est une norme informatique appliquée à la documentation, dont le but est d'assurer l'interopérabilité des systèmes (ISO/IEC, 2002, p. iii, [29]) :

« This International Standard provides a standardised notation for interchanchably representing information about the structure of information resources used to define topics, and the relationships between topics. A set of one or more interelated documents that employs the notation defined by this International Standard is called a ***topic map****. In general, the structural information conveyed by topic maps includes :*
— groupings of adressable information objects around topics ('occurrences'), and
— relationships between topics ('associations'). »

La première version (2000) décrivait en annexe l'architecture de la meta-DTD SGML basée sur la norme HyTime (ISO/IEC 10744 : 1997), appelée HyTM (pour HyTime Topic Maps).

La seconde version (2002) ajoute une deuxième architecture, orientée Web, une DTD XML, appelée XTM (pour XML Topic Maps).

Les « Topics » sont des sujets interprétables par la machine. Lors de sa création, le Topic, qui représente un sujet, devient un « objet » interprétable par la machine (Rath, 2003, p. 11, [30]). Les « Topics » ont trois caractéristiques principales :

⇒ noms (« base names ») : plusieurs synonymes intra et inter-lingistiques
⇒ occurrences (« occurrences ») : pointent vers les ressources (documents)
⇒ rôles (« playing roles in associations ») : relations entre les « topics »

Les types de relations gérées sont :

- hiérarchique (classe – subclasse)
- instanciation (classe – instance)
- associations (à choisir librement en fonction du domaine : ex. « is located in », « belongs to »)

Quelques éléments de la structure des Topic Maps (ISO/IEC, 2002, [29] ; Pepper, 2000, [31] ; Mondeca, [28] ; Rath, 2003, [30]) sont rappelés ici :

- « Topics », pouvant avoir plusieurs « Topic name » : les « Topics » s'organisent en classes selon la relation hiérarchique (« subclass » – « superclass ») et la relation d'instanciation (« class » – « instance »).
- « Occurrences », pointeurs qui relient les « Topics » aux documents et jouer des « Occurence roles » différents et « Occurrence role types » différents
- « Topic associations », relations étiquetées (« Association types »), hiérarchiques ou non, « Association roles »
- « subject identity » et « subject indicators »
- « facets », propriétés assignés aux ressources et « facet values », la valeur à assigner
- « topic characteristics »
- « scope », « themes »

1.2.1.4 ONTOLOGIES

Le terme est emprunté à la philosophie où il désigne la partie de la métaphysique qui étudie l'être. En intelligence artificielle, dans les systèmes à base de connaissances, ce qui existe est ce qui peut être représenté et l'univers du discours est construit par le set d'objets formalisés.

Dans le domaine de l'ingénierie des connaissances, le développement des ontologies occupe une place importante. Les pratiques issues des systèmes experts sont « révolues », il ne s'agit plus d'effectuer la modélisation psychologique des connaissances d'un expert, mais de la construction coopérative d'un modèle des connaissances (Charlet, 2000, p. 3, [32]), d'où le syntagme « système à base de connaissances ». L'ingénierie des connaissances est définie (Charlet, 2000, p. 2, [32]) comme :

> *« [...] l'étude de concepts, méthodes et techniques permettant de modéliser et/ou acquérir les connaissances pour les systèmes réalisant ou aidant les humains à réaliser des tâches se formalisant a priori peu ou pas. »*

Et les systèmes à base de connaissances (SBC ou KBS — Knowledge Based Systems) sont définis (Charlet, 2000, p. 5, [32]) ainsi :

> *« [...] est un système à base de connaissances tout système de manipulation d'inscriptions, pour peu que cette manipulation soit appréhendée et explicitement modélisée en fonction des usages et des fonctions qu'elle autorise. »*

La définition des ontologies, donnée par Thomas R. Gruber en 1993 (Gruber, 1993, [33] ; Gruber, 1995, p. 1, [34]) est la suivante :

> *« An ontology is an explicit specification of a conceptualization. »*

La « conceptualisation » signifie la construction d'un modèle abstrait qui représente un phénomène en ayant identifié les concepts pertinents, « explicite » signifie défini explicitement en termes informatiques, comme formalisation lisible par la machine (« machine readable »), « partagée », car l'ontologie devrait capter la connaissance unanimement partagée, consensuelle. (Ding, 2002a, p. 123, [35]). Les ontologies peuvent être générales ou dédiées à un domaine spécifique (Vickery, 1997, p. 279, [36]).

Une ontologie sera donc un vocabulaire partagé qui exprime la connaissance d'un domaine, définie par des classes d'objets, relations, fonctions, axiomes et instanciations. (Noy, 2001, p. 3, [37]) :

> *« For the purposes of this guide an **ontology** is a formal explicit description of concepts in a domain of discourse (**classes** (somtimes called **concepts**)), properties of each concept describing various features and attributes of the concept (**slots** (sometimes called **roles** or **properties**)), and restrictions of slots (facets (sometimes called role restrictions)). An ontology together with a set of individual **instances** of classes constitutes a **knowledge base**. »*

Selon B. Bachimont, la définition d'une ontologie est plutôt liée à sa construction (Charlet, 2000, p. 307, [32], article de B. Bachimont ; Bachimont, 1995, [38]) :

> *« [...] définir une ontologie pour la représentation des connaissances, c'est définir, pour un domaine et un problème donnés, la signature fonctionnelle et relationnelle d'un langage formel de représentation et la sémantique associée. »*

Construire une ontologie peut avoir pour but (Noy, 2001, p. 1, [37]) :

- ⇒ partager la connaissance et le vocabulaire utilisé, même dans le contexte du traitement par la machine
- ⇒ pouvoir réutiliser le résultat
- ⇒ rendre explicite le domaine de connaissance
- ⇒ séparer la connaissance du domaine de la connaissance opérationnelle
- ⇒ analyser le domaine de connaissance

Une fois construites (Gandon, 2002, [39]), selon des méthodologies et dans des domaines divers, comme, par exemple, le médical ou l'audio-visuel (Zweigenbaum, 1996, [40] ; Dechilly, 2000, [41]) la maintenance et la mise à jour des ontologies (« evolving ») est à envisager sous divers aspects (Ding, 2002b, [42]).

Les relations gérées par les ontologies peuvent être de type (voir aussi partie 1.2.3 DÉMARCHES DE CONSTRUCTION, page 28, plus particulièrement partie 1.2.3.2 HIÉRARCHISATION DES CONCEPTS, page 32) :

- ⇒ hiérarchique : « is a kind of » (classes – subclasses)
- ⇒ instanciation : « is a »
- ⇒ partitive : « is a part of »

Et aussi, des relations étiquetées librement selon les besoins du domaine :

- ⇒ cause
- ⇒ fait par
- ⇒ ...

Du point de vue de l'expression informatique, différents langages peuvent être utilisés lors de la mise en place des ontologies. Pour l'instant, il n'y a pas de norme concernant l'interopérabilité des différents langages utilisés, mais quelques standards sont recommandés par le consortium W3C. Parmi les langages utilisés pour la réalisation des ontologies, on peut citer :

- ⇒ RDF(S) (= Resource Description Framework (RDF) + RDF Schema (RDFS))
- ⇒ OIL (Ontology Inference Layer)
- ⇒ DAML (Darpa Agent Markup Language)
- ⇒ OWL (Ontology Web Language) (= DAML + OIL)

Et aussi :

- ⇒ UML: Unified Modeling Language
- ⇒ KIF : Knowledge Interchange Format

Quelques éléments constitutifs des ontologies sont rappelés ici (Noy, 2001, [37]) :

- classes (« classes »), appelées quelquefois concepts (« concepts ») : les classes décrivent les concepts du domaine, organisés en hiérarchies superclasses-subclasses (« superclass » - « subclass ») ; les subclasses représentent des concepts plus spécifiques
- « slots », appelés aussi rôles (« roles ») ou propriétés (« properties ») : les « slots » décrivent les propriétés des classes
- « facets », ou « roles restrictions » : décrivent les valeurs attribuées aux « slots »
- instanciations (« instances ») : une ontologie et un set d'instanciations forment une base de connaissances (Knowledge Base)

1.2.2 COMPARATIF DES TYPES DE RELATIONS GÉRÉES

Le comparatif des quatre systèmes d'organisation des connaissances selon les types de relations gérées peut être représenté de façon schématique (voir Figure 3 - Relations gérées par les taxonomies / thésaurus / Topic Maps / ontologies, page 27 et annexe page 148) :

La relation hiérarchique, de type « est une sorte de » (« is a kind of ») est commune à tous ces systèmes et correspond à la façon dont les humains classent les objets (conceptuels ou du monde réel), par inclusion de classes.

La relation partitive « est une partie de » (« is a part of ») fait aussi partie des relations gérées par tous ces systèmes, mais, dans le cas de la taxonomie et du thésaurus, elle n'est pas différenciée de la relation hiérarchique. Selon les possibilités citées dans les normes des thésaurus, elle peut recevoir une notation particulière.

L'instanciation « est un » (« is a ») est prise en compte, de façon différente toutefois : elle est indifférenciée de la relation hiérarchique (inclusion de classes) dans les taxonomies et les thésaurus. Dans le cas des Topic Maps et des ontologies (et, quelquefois des taxonomies mises en place dans les moteurs de recherche, qui s'appuient sur des réseaux de concepts), l'instanciation est une relation différente de la relation hiérarchique.

La synonymie est gérée à partir du thésaurus.

La relation appelée associative et les facettes des thésaurus sont détaillées en relations étiquetées dans les Topic Maps et les ontologies, comme, par exemple : cause (« a pour cause »), conséquence (« a pour conséquence »), matériau (« fait en »), produit (« fait par ») etc. Le nom des relations est à choisir librement, mais à définir formellement de façon stricte, selon des propriétés possédées en commun par les concepts.

Représentation schématique des types de relations gérées par les systèmes d'organisation des connaissances (KOS)

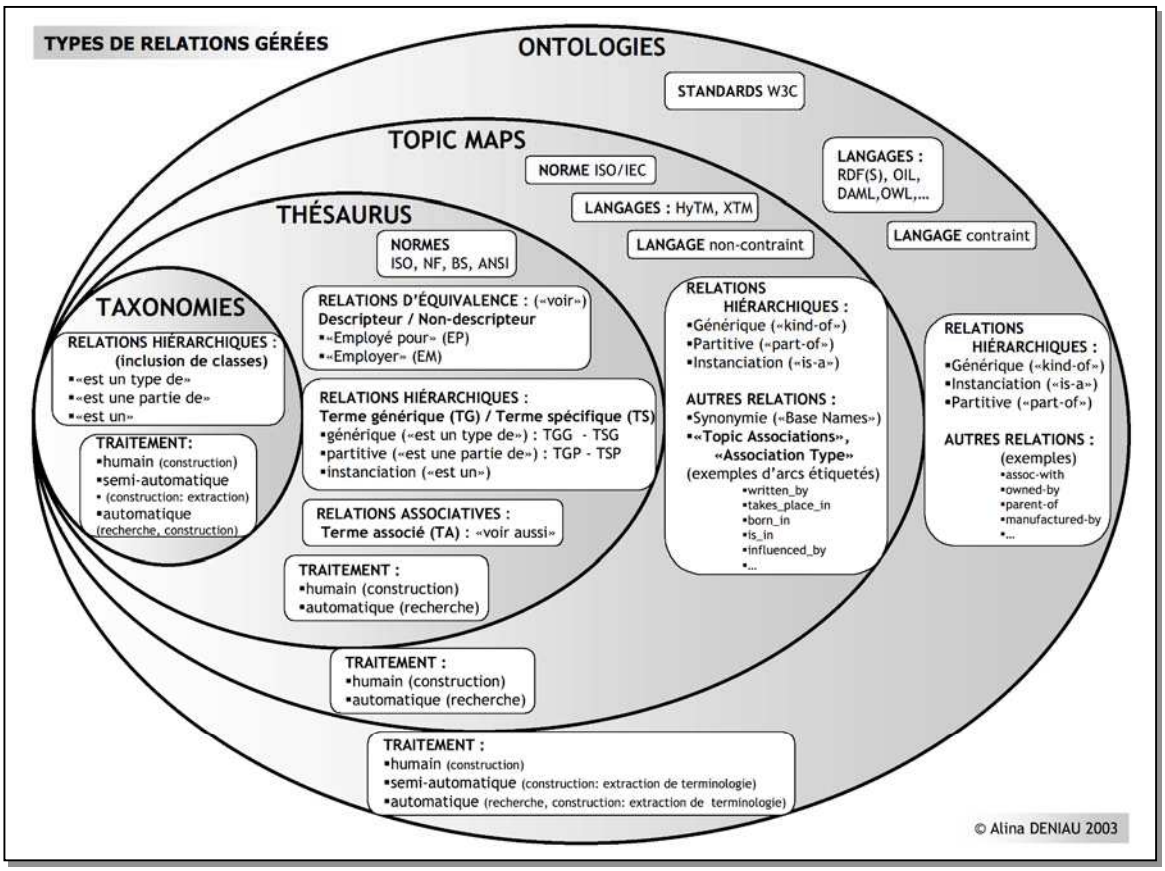


Figure 3 - Relations gérées par les taxonomies / thésaurus / Topic Maps / ontologies

1.2.3 DÉMARCHES DE CONSTRUCTION

Concernant la méthodologie employée dans la construction de systèmes d'organisation des connaissances, on peut distinguer plusieurs phases. Sont laissées de côté ici les phases d'analyse des besoins et l'estimation des possibilités de réutilisation des systèmes existants, qui font partie de la démarche globale du projet.

Sont laissées de côté aussi les phases liées à la mise en place du système hiérarchisé de concepts en fonction des contraintes informatiques. Le constat du mélange de ces phases est valable pour la littérature concernant la construction d'ontologies, qui ne fait pas tellement la différence entre l'objet intellectuel et sa mise en oeuvre informatique.

Les phases qui seront discutées dans cette partie sont :

- collecte des termes
- hiérarchisation des concepts (organisation des relations)

Toutes les étapes de travail lors de la construction des THÉSAURUS sont décrites de façon très détaillée par Michèle Hudon (Hudon, 1994, [27]).

Quant aux nouvelles structures comme les ONTOLOGIES et les TOPIC MAPS, la distinction de ces deux phases de travail n'est pas toujours faite dans leur formalisation intellectuelle. Mais ces systèmes sont très formalisées du point de vue informatique : pour les ontologies, par exemple, des phases de définition des propriétés des concepts, de leur structure interne (« internal structure of concepts ») — « classes », « slots », « facets », « instances » — suivent la phase de hiérarchisation (Noy, 2001, [37]).

Si chez certains auteurs, la distinction entre la construction intellectuelle et le formalisme informatique n'est pas claire, des auteurs comme B. Bachimont (Charlet, 2000, pp. 305-323, [32]) ou B. Biébow et S. Szulman (Charlet, 2000, pp. 325-336, [32]) distinguent plus finement les phases de conception de la structure hiérarchique (et le treillis[6]) de l'ontologie (concepts et relations).

[6] Le treillis (« lattice » en anglais) est un réseau de concepts où toutes les relations sont étiquetées.

Selon B. Bachimont une ontologie se caractérise selon trois niveaux (Bachimont, 1995, [38]) ; Charlet, 2000, p. 322, [32], article de B. Bachimont) :

- Le niveau sémantique ou interprétatif (« ontologie régionale ») : l'ontologie est un arbre de concepts sémantiques.
- Le niveau formel ou référentiel (« ontologie référentielle ») : l'ontologie est un treillis de concepts formels. Un treillis (« lattice » en anglais) est un réseau de concepts où toutes les relations sont étiquetées.
- Le niveau opérationnel ou computationnel (« ontologie computationnelle ») : l'ontologie est un treillis computationnel. Le treillis reçoit une expression compréhensible par la machine.

Selon B. Biébow et S. Szulman (Charlet, 2000, pp. 332-333, [32]) :

> *« La première étape consiste donc à établir une liste de termes à partir des textes, ce qui nécessite au préalable de constituer un corpus pertinent décrivant le domaine de l'application. [...] L'étape suivante mène à la conceptualisation de chaque terme. L'ingénieur de la connaissance analyse tous les usages du terme dans le corpus afin de déterminer les différents sens (notions) dénotés par le terme et donne une définition en langage naturel de chaque notion. Lors de la troisième étape, cette définition doit être traduite dans un formalisme [...]. »*

Concernant les TOPIC MAPS, la méthode de construction, a effectuer avant la formalisation informatique serait : à partir d'un brainstorming, d'abord, identifier les sujets, futurs « topics », ensuite établir les synonymies dans la liste et identifier les classes et les relations, enfin, identifier les « scope » (pour permettre, par exemple, le filtrage en fonction des utilisateurs) (Rath, 2003, [30]) :

> *«As brainstorming is – with intention – a creative but quite disorganised process, the identified list of subjects has to be organised before it becomes input to the design phase.*
>
> *1. Ensure that all your listed subjects are really different subjects and not just synonyms. If you find synomyms just record them as alternative names for a subject.*
>
> *2. Look at the listed subjects and identify the classes for topics, occurrences, and associations, as well as the associations roles. You might also find instances of topics, occurrences, or associations in your list. Note them as examples for the appropriate classes.*
>
> *3. List subjects, which might be used in scope to define views, context informations, or access rights. »*

La collecte des termes et la hiérarchisation des concepts sont, donc, des étapes obligatoires pour la construction de toute structure qui s'appuie sur un réseau sémantique.

1.2.3.1 COLLECTE DES TERMES

Concernant la collecte des termes, les spécialistes de la construction de THÉSAURUS (Hudon, 1994, p. 77, [27]), distinguent deux démarches possibles : analytique et synthétique, et une troisième, combinaison des deux, à adapter en fonction des réalités du terrain.

- Démarche analytique (dite aussi déductive)

Il s'agit de collecter les termes en consultant la littérature de spécialité, notamment les articles de périodiques (dont la qualité est reconnue par les spécialistes), les actes de congrès, rapports de recherche. Il est conseillé de tenir compte, autant que possible, des questions des usagers (en les indexant). Les termes collectés doivent être validés par des experts du domaine, de préférence au moyen de la communication directe (ou en créant un comité consultatif d'experts) plutôt que par l'intermédiaire de documents écrits.

- Démarche synthétique (dite aussi inductive, globale ou philosophique)

Les termes seront collectés dans les ouvrages de référence « qui décrivent la structure conceptuelle d'un domaine et qui en ont déjà fait un inventaire terminologique » (Hudon, 1994, p. 77, [27]). Il s'agit de : dictionnaires et encyclopédies spécialisées, manuels utilisés dans l'enseignement de la discipline, classifications et nomenclatures, listes de vedettes-matières et thésaurus dans des domaines connexes, tables de matières et index, banques de terminologie, lexiques et glossaires.

- Démarche combinée

Dans la pratique, la combinaison des deux méthodes est fonction du domaine à décrire (champ « jeune » ou bien établi) et des conditions matérielles : temps de réalisation, nombre de personnes, etc.

Lors de la réalisation d'un thésaurus, il est conseillé d'utiliser un tiers du temps du projet pour réaliser la collecte des termes candidats-descripteurs.

Concernant la construction des ONTOLOGIES, certaines sources considérent comme utile (« useful ») d'avoir une liste initiale de termes (Noy, 2001, p. 6, [37]), mais sans offrir plus de précisions concernant la façon de l'obtenir :

> *« Initially, it is important to get a comprehensive list of termes without worrying about overlap between concepts they represent, relations among terme, or any properties that the concept may have, or whether the concepts were classes or slots. »*

Selon d'autres sources (Vickery, 1997, p. 282, [36], citant un article de 1996 de Uschold et Gruninger[7]), il est recommandé de consulter la littérature du domaine et de recourir à une session de brainstorming avec les experts pour déterminer quels sont les termes - concepts importants (« concept terms ») :

> *« Concept terms are collected by scanning the litterature of the domain, and by consulting domain experts. A 'brainstorming' session with experts is recommended, to produce significant terms and their relative importance. »*

B. Bachimont (Charlet, 2000, pp. 308-309, [32]), fait remarquer que la connaissance dont on dispose doit être modélisée (afin de « résoudre un problème » [8]), « à partir de l'expression linguistique des connaissances », trouvée dans les documents sélectionnés constitués en corpus :

> *« Le travail de modélisation doit s'effectuer à partir de documents attestés dans la pratique d'un domaine et rassemblés en un corpus. Le corpus est constitué de documents produits dans le contexte où le problème à résoudre se pose. Ce sont, par exemple, des documentations techniques, des ouvrages de référence, des documents de travail, des manuels propres au domaine ou à l'industrie concernée, ou bien encore la transcription d'interviews menées avec les spécialistes. [...] Le corpus comporte l'expression des notions à modéliser. »*

Selon B. Biébow et S. Szulman (Charlet, 2000, p. 332, [32]), la première étape du processus de construction d'une ontologie est la constitution d'une liste de candidats termes issus d'un corpus du domaine, obtenus par extraction automatique :

> *« La première étape consiste donc à établir une liste de termes à partir des textes, ce qui nécessite au préalable de constituer un corpus pertinent décrivant le domaine de l'application. À partir de ce corpus, un outil d'extraction fournit une liste de candidats termes, parmi lesquels l'ingénieur de la connaissance sélectionne un ensemble de termes à l'aide d'un expert du domaine. »*

Concernant la participation des experts à la construction de l'ontologie, B. Bachimont affirme la différence de point de vue entre celui des systèmes experts (« pratiques aujourd'hui révolues », selon l'introduction de l'ouvrage (Charlet, 2000, p. 3, [32]) et le travail de « l'ingénieur de la connaissance » (Charlet, 2000, p. 311, [32], article rédigé par B. Bachimont) :

> *« [...] l'expert [...] ne délivre pas ses connaissances propres, mais ce que signifient, selon lui, les expressions linguistiques du corpus. Par ailleurs, l'ingénieur de la connaissance n'est pas là pour accoucher l'expert des connaissances dont il n'a pas forcément conscience, mais pour structurer les expressions linguistiques et les mettre en forme selon une méthodologie pour que, normalisées, elles puissent tenir lieu d'expression de concepts. »*

[7] Pour la démarche de collecte des termes, B. Vickery cite l'article de M. Uschold et M. Gruniger « Ontologies : Principles, Methods and Applications » publié en 1996 dans la revue Knowledge Engeneering Review.

[8] Vision dans laquelle on reconnaît les avatars des systèmes experts dans l'ingénierie des connaissances.

1.2.3.2 HIÉRARCHISATION DES CONCEPTS

L'organisation des termes et la détermination des relations qui les unissent est considérée comme la partie la plus difficile.

Dans le cas des THÉSAURUS, cette phase obéit à des critères très explicites, formalisés depuis longtemps (première version de la norme ISO 2788 date de 1974).

Plusieurs méthodes de travail sont considérées possibles pour aboutir à la hiérarchisation des concepts :

- « bottom-up »
- « top-down »
- combinée (mixte ou « centrifuge »)

Les termes viennent de l'ingénierie des connaissances (Charlet, 2000, [32]). Ces méthodes seront présentées ici, selon la littérature concernant la construction de thésaurus et ontologies.

1.2.3.2.1 « BOTTOM-UP » OU DÉMARCHE THÉSAURUS

C'est la démarche de construction de thésaurus. On part de la liste à plat des candidats-descripteurs et on construit les relations en regroupant les termes dans des classes de plus en fonction des regroupements spécifiques au domaine et à la structure (entreprise ou service) à qui le thésaurus est destiné.

Le terme « bottom-up », utilisé par M. Uschold et M. Gruniger en 1996, dans un article décrivant la démarche de construction des ONTOLOGIES est traduit en français par « démarche ascendante » (Charlet, 2000, pp. 26-27 et p. 293, [32] ; Gandon, 2002, p. 425, [39]). La démarche « bottom-up » pour définir les classes et leur hiérarchie est décrite succinctement (Noy[9], 2001, p. 7, [37]) ainsi :

> *« A **bottom-up** developpment starts with the definition of the most specific classes, the leaves of the hierarchy, with subsequent grouping of these classes into more general concept. »*

Concernant la construction des ontologies, d'autres sources (Charlet, 2000, pp. 26-27, [32], article de M. Dubrieux-Coquebert et B. Houriez) rapprochent la démarche ascendante des méthodes de modélisation de type KOD (« Knowledge Oriented Design »). D'autres auteurs précisent que dans la démarche « bottom-up », qu'ils appellent sémasiologique (Charlet, 2000, pp. 292-394, [32], article de P. Frath, R. Oueslati et F. Rouselot), on commence par « faire l'inventaire des connaissances du texte pour ensuite construire l'ontologie ».

[9] L'article de N. Noy et D. L. McGuinness de l'Université de Stanford cite le papier publié par M. Uschold et M. Gruniger « Ontologies : Principles, Methods and Applications » en 1996 dans la revue Knowledge Engeneering Review.

La méthode appliquée lors de la conception de THÉSAURUS est la description la plus explicite de la méthode de travail qui consiste à partir d'une liste de termes sélectionnés, choisir les concepts et hiérarchiser les relations (donc, « bottom-up »). Elle est détaillée dans les ouvrages traitant de la construction des thésaurus, comme celui de Michèle Hudon[10] (Hudon, 1994, p. 77, [27]).

La trame de cette méthode sera donc utilisée afin de présenter l'enchaînement des phases de construction. Une partie des actions à mener et des vérifications et validations à effectuer afin de hiérarchiser les concepts figurent, de façon assez éparse, dans l'article de N. Noy et D. L. McGuinness (Noy, 2001, [37]). Des remarques formulées par B. Biébow et S. Szulman (Charlet, 2000, p. 331, [32]) ou par B. Bachimont concernant l'héritage de propriétés vont dans le même sens (Bachimont 1995, [38] ; Charlet, 2000, p. 305-323 [32], partie rédigée par B. Bachimont) et sont intégrées dans la présentation.

❑ Premiers groupements sémantiques (thésaurus)

Lors de la construction d'un THÉSAURUS, afin de faciliter la manipulation des termes, ils seront regroupés en classes d'une centaine de termes au maximum. Le groupement peut s'effectuer :

- ⇒ par thèmes : à déterminer en fonction du domaine
- ⇒ par facettes : processus, outil, propriété, phénomène, équipement, être vivant, matériau
- ⇒ combinaison des deux : par thèmes d'abord complétés par de facettes ensuite

Cette phase n'est pas présentée dans la littérature concernant la construction d'ontologies.

❑ Choix des descripteurs (thésaurus) / Choix du nom de la classe ou du concept (ontologie)

La normalisation des termes d'un THÉSAURUS est explicite : la forme des descripteurs (en français) est, généralement, un nom au masculin singulier. Les normes sont explicites aussi quant aux choix concernant les mots composés et les cas d'acceptation des autres classes grammaticales comme descripteurs.

Pour choisir les descripteurs dans un thésaurus, il s'agit d'abord de déterminer :

- ⇒ synonymes
- ⇒ quasi-synonymes
- ⇒ antonymes

Le choix du terme préférentiel (descripteur) se fait surtout en fonction de l'usage le plus répandu. Le terme le plus connu sera choisi comme descripteur dans un thésaurus.

[10] La Canadienne Michèle Hudon est un des spécialistes reconnus des thesaurus.

Dans la construction des ONTOLOGIES, quelques explications sont donnés afin d'orienter le choix du nom des classes (Noy, 2001, pp. 12-13, [37]). Pour l'usage du singulier ou du pluriel, le conseil sera de faire un choix et de s'y tenir (Noy, 2001, p. 21, [37]) :

« Classes represents concepts in the domain and not the words that denote. »

« Synonyms for the same concept do not represent different classes. »

« Define a naming convention for classes and slots and adhere to it. »

❑ Reconstitution des hiérarchies

Si, dans la construction du THÉSAURUS, les premiers regroupements ont été faits sur des critères explicites, la hiérarchisation peur être faite assez vite.

⇒ VALIDATION DES RELATIONS GÉNÉRIQUES

Dans la construction des THÉSAURUS, la validation des relations hiérarchiques (Terme générique - Terme spécifique) se fait selon la règle « certains/tous » : « Certains A sont des B. Tous les B sont des A. » (Ex. : Tous les chats sont des félins. Certains félins sont des chats.) (Hudon, 1994, pp. 99-102, [27]).

Concernant la vérification des relations hiérarchiques dans les ONTOLOGIES, il s'agit d'une relation de type « is-a » ou « kind-of » (Noy, 2001, p. 12, [37]) et le niveau de généralité des concepts de même niveau doit être équivalent (Noy, 2001, p. 14, [37]) :

« A subclass of a class represents a concept that is a « kind of » the concept. »

« All the siblings in the hierarchy (except for the ones at the root) must be at the same level of generality. »

La vérification de la transitivité de la relation hiérarchique dans les ontologies est exprimée ainsi (Noy, 2001, p. 13, [37]) :

« If B is a subclass of A and C is a subclass of B, then C is a subclass of A. »

B. Biébow et S. Szulman, pour qui la « conceptualisation » est la deuxième étape de la construction des ONTOLOGIES (Charlet, 2000, p. 331, [32]), décrivent ainsi la « conceptualisation bottom-up » (« les concepts de regroupement ») :

« Les concepts qui partagent des propriétés communes peuvent être regroupés sous un concept qui les subsume. »

L'héritage de propriétés dans la construction des ontologies est considéré comme un principe fondamental par B. Bachimont, qui, dès 1995, établit quatre grands principes à respecter pour déterminer la place d'un concept dans une ontologie (Bachimont, 1995, pp. 78-79, [38] ; Charlet, 2000, p. 313, [32], article de B. Bachimont) :

- Principe de communauté avec le père : nécessité de déterminer l'identité avec le père (la caractéristique commune)
- Principe de différence avec le père : nécessité de préciser la différence avec le père
- Principe de différence avec les frères : nécessité de préciser la différence avec les frères
- Principe de communauté avec les frères : nécessité de préciser la dimension sémantique commune

Ces quatre principes forment une grille de lecture qui force la définition de chaque principe fondée sur des traits différentiels, et devrait permettre de construire l'ontologie de façon organisée et méthodique, « de manière systématique », le but étant d'obtenir une ontologie réutilisable (Charlet, 2000, p. 321, [32], article de B. Bachimont). La méthode décrite par B. Bachimont a été utilisée pour la construction de l'ontologie Menelas, dans le domaine médical (Zweigenbaum, 1996, [40]), et une ontologie est en cours de construction dans le domaine audio-visuel (Dechilly, 2000, [41]).

Concernant la gestion de la polyhiérarchie, dans la construction des THÉSAURUS, il est conseillé de la limiter aux cas absolument indispensables et de vérifier les niveaux hiérarchiques (grand-père — père — fils) (Hudon, 1994, p. 102, [27]).

Dans les ONTOLOGIES, la polyhiérarchie, acceptée par la plupart des systèmes, s'appelle « héritage multiple » (« multiple inheritance »).

⇒ VALIDATION DES RELATIONS ASSOCIATIVES

Les types de relations qui sont acceptées comme relations associatives dans les THÉSAURUS sont présentées par la norme : cause - son effet, tout - partie, action - son agent, action - son produit, action - son objet, action - lieu de l'action, science - son objet, objet - sa propriété, objet - son application, matériau - produit (voir aussi partie 1.2.1.2 THÉSAURUS, page 18).

Dans les ONTOLOGIES, ces types de relations sont étiquetées et sont prises en charge par la définition des propriétés des concepts.

La méthode de travail décrite pour concevoir les thésaurus, extrêmement formalisée, peut être appliquée (en l'adaptant) aux autres types de systèmes d'organisation des connaissances (KOS) : on obtient ainsi la validation de toutes les relations, vérifiées à l'aide de critères explicites et logiques.

1.2.3.2.2 « TOP-DOWN » OU DÉMARCHE CLASSIFICATOIRE

Citée dans la littérature concernant le développement des ontologies, la démarche « top-down » est une démarche de type classificatoire. La traduction française est « approche descendante » (Charlet, 2000, pp. 26-27 et p. 293, [32] ; Gandon, 2002, p. 425, [39]). L'approche est décrite ainsi (Noy, 2001, p. 6, [37]) :

> *« A **top-down** developpment process starts with the definition of the most general concepts in the domain and subsequent specialization of the concepts. »*

On part d'un niveau très général et on crée les subclasses, de plus en plus spécialisées.

D'autres sources, visiblement proches des systèmes experts, décrivent l'approche descendante comme étant une méthode « qui se focalise rapidement sur la formulation d'un modèle d'expertise » (Charlet, 2000, pp. 26-27, [32], article de M. Dubrieux-Coquebert et B. Houriez) et la rapprochent de la méthode de modélisation KADS (« Knowledge Acquisition and Design Structuring »).

Selon d'autres auteurs, l'approche « top-down », qu'ils appellent onomasiologique (Charlet, 2000, pp. 292-394, [32], article de P. Frath, R. Oueslati et F. Rouselot), « va du niveau conceptuel vers le texte », mais son point faible « est la phase d'acquisition, largement manuelle ». Ils conseillent donc de la faire précéder d'une approche « bottom-up » pour la phase d'acquisition (collecte des termes) :

> *« Pour cette approche, on commence par poser des ensembles de concepts et de relations couvrant idéalement tout le domaine. [...] l'analyse manuelle des textes permet ensuite de remplir la structure abstraite et de construire l'ontologie. »*

Les conseils concernant la vérification des hiérarchies (Noy, 2001, [37] ; Bachimont, 1995, [38] ; Charlet, 2000, pp. 305-323, [32], article de B. Bachimont) développées dans la partie 1.2.3.2.1 « TOP-DOWN » OU DÉMARCHE CLASSIFICATOIRE s'appliquent à cette démarche aussi, mais l'enchaînement des actions à mener n'étant pas spécifié, ils ne sont pas rappelés ici.

Pour B. Biébow et S. Szulman, la « conceptualisation » est la deuxième étape de la construction des ontologies (Charlet, 2000, p. 331, [32]), et la « conceptualisation « top-down » » (« les concepts de structuration de haut niveau ») est vue ainsi :

> *« Définis dans les niveaux les plus élevés de l'ontologie, ces concepts servent à structurer la pensée du concepteur et sont généralement liés à l'application ou au domaine et à ses activités. »*

L'article émet l'appréciation suivante (Charlet, 2000, p. 332, [32], article rédigé par B. Biébow et S. Szulman) :

> *« Ce type de conceptualisation permet de découper très vite la base en sous-parties plus faciles à appréhender, mais ne correspond pas forcément à une réelle nécessité, il est ensuite très difficile de remettre ce découpage en question, car cela remet en question tous les niveaux inférieurs de la base. »*

1.2.3.2.3 DÉMARCHE COMBINÉE

Dans le développement des ONTOLOGIES, la démarche « combinée » (« combination development process ») est appelée par F. Gandon approche « centrifuge » (Gandon, 2002, p. 425, [39]) et par d'autres sources « approche mixte » (Charlet, 2000, pp. 26-27, [32], article de M. Dubrieux-Coquebert et B. Houriez). Cette démarche (Noy, 2001, p. 7, [37]) est définie ainsi :

> *« A combination development process is a combination of the top-down and bottom-up approaches : We define the more salient concepts first and then generalize and specialize them appropriately. »*

Il s'agit donc de créer des concepts en commençant par les concepts du milieu de la hiérarchie (« middle-level concepts ») et de créer ensuite des classes supérieures (« a few top-level concepts ») et des concepts plus spécifiques (« a few specific concepts »), de façon « appropriée ».

M. Dubrieux-Coquebert et B. Houriez (Charlet, 2000, pp. 29-31, [32]), développent les raisons de choisir une approche mixte, mais leur raisonnement semble fondé sur le fait que ni la méthode KOD (Knowledge Object Design), ni la méthode KADS (Knowledge Acquisition and Design Structuring) ne correspondent aux nécessités imposées par la construction des ontologies.

De même que pour la démarche « top-down », les conseils concernant la vérification des hiérarchies (Noy, 2001, [37] ; Bachimont, 1995, [38]) s'appliquent (voir partie 1.2.3.2.1 « BOTTOM-UP » OU DÉMARCHE THÉSAURUS, page 32), mais l'enchaînement des actions à mener n'étant pas spécifié, ces conseils ne seront pas rappelés à nouveau.

Cette méthode « centrifuge », « combinée » ou « mixte », consistant à combiner des phases inspirées par la démarche ascendante avec des phases inspirées par la démarche descendante, a été préférée par F. Gandon lors de la réalisation du projet O'CoMMA, démarche décrite dans sa thèse (Gandon, 2002, [39]).

1.2.3.3 CONCLUSIONS

Lors de la mise en place d'un système d'organisation des connaissances dans un moteur de recherche, que ce soit une taxonomie ou un thésaurus, (référentiel d'entreprise), la nécessité d'adopter une méthode de travail se fera sentir.

La méthode dite « bottom-up » ou ascendante est la plus explicitement formalisée des trois et son application pendant les 40 dernières années lors de la conception de thésaurus a fait ses preuves. Son manque de formalisme informatique n'est pas un frein, car, s'agissant d'obtenir une construction intellectuelle (sémantique), sa formalisation informatique (computationnelle) peut être vue comme l'étape suivante.

1.3 LES MOTEURS DE RECHERCHE ET LEUR FONCTIONNEMENT

Après les avoir défini, pour mieux comprendre le fonctionnement de ces logiciels, un aperçu des fondements théoriques des disciplines qui participent à leur réalisation sera suivi d'une cartographie des acteurs présents sur le marché des moteurs de recherche et d'une description fonctionnelle illustrée d'exemples d'un produit particulier : K2 Enterprise de Verity.

❑ Définition des moteurs de recherche

Une définition fonctionnelle des moteurs de recherche a été fournie par la consultante Catherine Leloup (Leloup, 1998, p. 17, [43]) :

> *« C'est un outil qui permet d'extraire d'une information, principalement textuelle, les mots ou termes qui la représentent le mieux et de les stocker dans un index : le même outil parcourt ensuite cet index afin d'identifier les termes les plus pertinents par rapport à ceux de la question de l'utilisateur, puis de trier les informations à lui fournir en retour. »*

Mais, précise la suite :

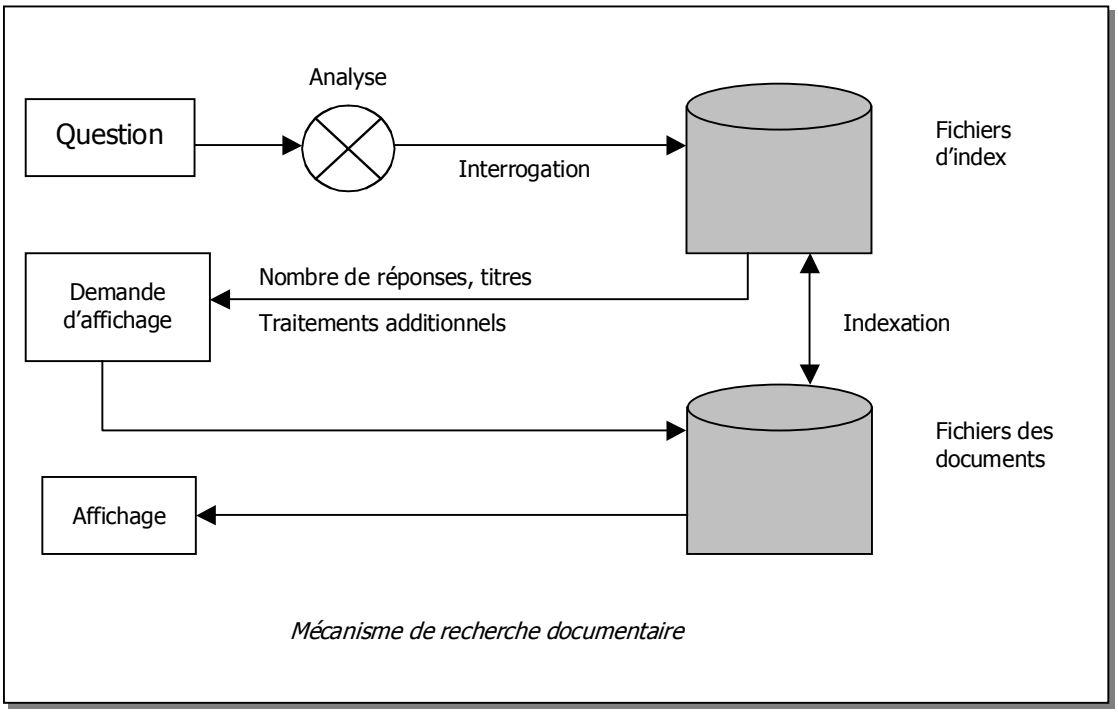
> *« Un moteur est utilisé pour construire une application ou un service, dont l'utilisateur final ne verra que l'interface d'interrogation [...]. »*

À la question posée par l'usager (voir Figure 4 - Mécanisme de recherche, page 39), le moteur de recherche doit pouvoir répondre en un temps suffisamment court (Leloup, 1998, p. 28, [43]) :

> *« En effet, on ne baptisera du nom de « moteur d'indexation et de recherche » que des outils capables d'explorer en un temps raisonnablement court, c'est-à-dire quelques secondes, des collections de milliers, voire de millions, de documents. Les autres iront jouer dans la cour des « utilitaires ». »* [11]

Afin d'obtenir des temps de réponse courts, les éditeurs de logiciels utilisent différents modes d'indexation. Une typologie de l'indexation pratiquée par les moteurs de recherche est proposée par Philippe Lefèvre (Lefèvre, 2000, p. 108-127, [4]).

[11] Les utilitaires fournis avec les systèmes d'exploitations permettent de faire diverses recherches portant sur le nom ou le contenu d'un fichier, mais les temps de réponse aux requêtes sont extrêmement longs, sans commune mesure avec ceux des moteurs de recherche.



Source : Catherine Leloup, Moteurs d'indexation et de recherche, Eyrolles, 1998, p. 29 (Leloup, 1998, [43])

Figure 4 - Mécanisme de recherche

1.3.1 MOTEURS DE RECHERCHE ET TRAITEMENT AUTOMATIQUE DES LANGUES « NATURELLES » (TALN)

Les moteurs de recherche commercialisés actuellement sont des applications industrielles de ce que l'on appelle aujourd'hui le TALN (en anglais NLP, « Natural Language Processing »). Le sigle est souvent traduit par « traitement automatique du langage naturel » ou « traitement automatique des langues naturelles ». Certains auteurs, et non des moindres (Fuchs, 1993, p. 8-10, [44] ; Rastier, 1994, p. 9, [45] ; Habert, 1997, p. 8, [46]), dénoncent le calque anglais et font valoir que les syntagmes « langue naturelle » et « langage naturel » sont inappropriés, car le français dispose de deux termes : « langue » et « langage » pour dissocier les deux notions, alors que l'anglais en possède un seul, « language », qui couvre les deux sens. D'où la nécessité (en anglais) de qualifier la langue de naturelle pour la différencier du langage, qui peut être artificiel, comme les langages informatiques.

Querelle terminologique mise à part, le sigle TALN s'est imposé et il est aujourd'hui incontournable, suivi de près par TAL (traitement automatique des langues, plus conforme à l'usage français).

1.3.1.1 DÉFINITION DU TALN

Ensemble de techniques qui s'appuient sur des théories linguistiques, le TALN (NLP) vise l'analyse et la compréhension du langage humain par la machine (Chowdhury, 2003, p. 51, [47]) :

> *« Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developped to make computer systems understand and manipulate natural language to perform desired task. »*

Selon les termes de Catherine Fuchs, (Fuchs, 1993, p. 13, [44]) :

> *« [...] **les traitements automatiques des langues** ont pour objet des **données linguistiques** (**textes**) exprimées dans une **langue** (« **naturelle** »), et [...] pour pouvoir traiter automatiquement ces données, il faut être capable d'expliciter les **règles de la langue**, de les représenter dans des **formalismes** opératoires et calculables et de les implémenter à l'aide de **programmes** ».* [12]

[12] C'est C. Fuchs qui souligne.

1.3.1.2 DISCIPLINES IMPLIQUÉES

Le traitement automatique des langues naturelles (TALN) fait intervenir de multiples domaines : la linguistique, l'informatique, les mathématiques, la statistique, la logique, les sciences cognitives (Mahmoudi, 1997, pp. 16-28, [48]). La modélisation du langage est considérée comme un des domaines essentiels de l'intelligence artificielle (IA). Les disciplines à la base du TALN (NLP) sont (Chowdhury, 2003, p. 51, [47]), donc :

> « *The foundation of NLP lie in a number of disciplines, namely, computer and information sciences, linguistics, mathemathics, electrical and electronic engineering, artificial intelligence and robotics, and psychology.* »

Seront rappelés ici quelques principes et méthodes employés en TALN, issus de disciplines comme : la linguistique, l'informatique, les mathématiques et la statistique, la logique.

1.3.1.2.1 LA LINGUISTIQUE

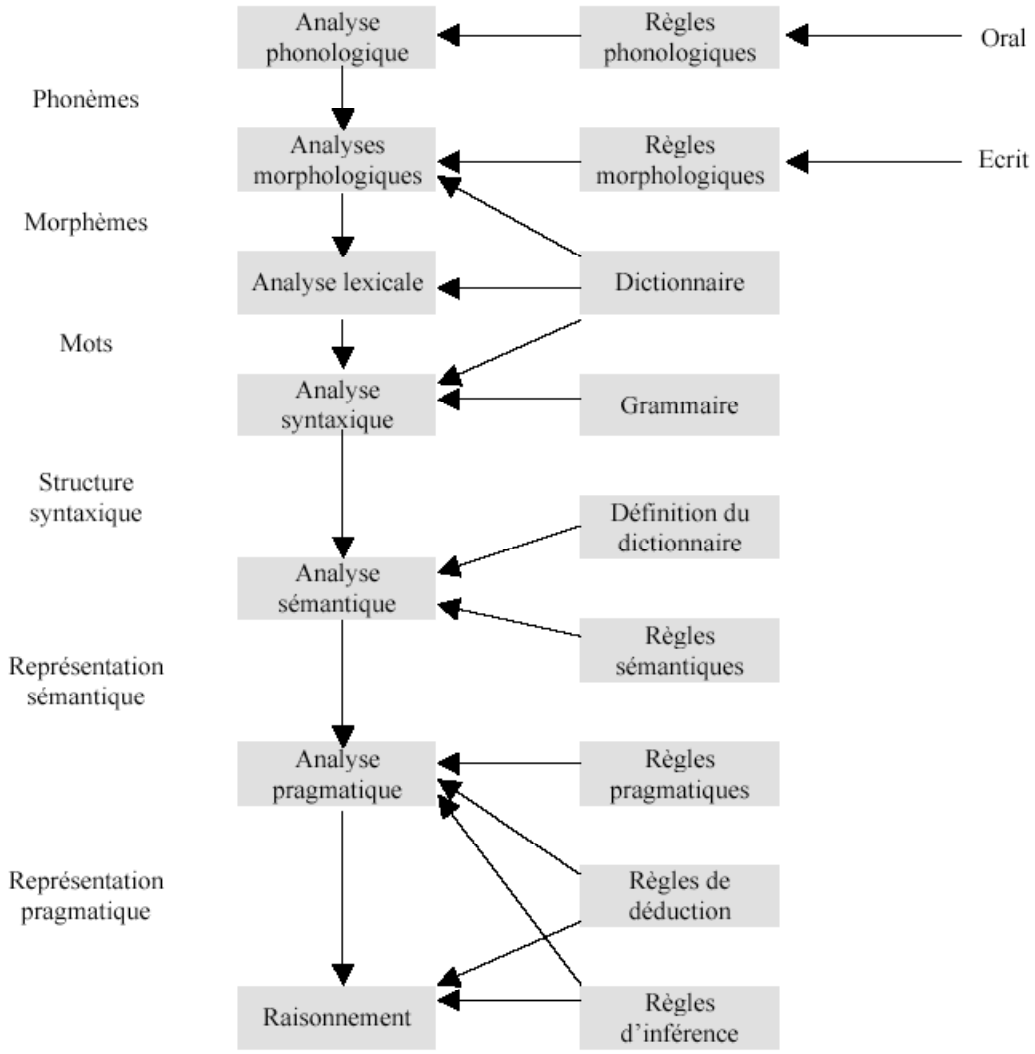
Les traitements automatiques appliqués aux textes s'appuient sur des modèles théoriques de la langue et de ses niveaux d'analyse. L'analyse de textes par ordinateur (ATO) (SATO, 2001, [49]) ; OTIL, 2003, [50]) se fonde sur une analyse linguistique approfondie préalable à toute tentative de formalisation informatique.

Dans le traitement automatique de la langue (TALN), les niveaux d'analyse linguistique — phonologique, morphologique, lexicale, syntaxique, sémantique, pragmatique et raisonnement (voir Figure 5 – Niveaux d'analyse de la langue (TALN), page 42) — définis par Gérard Sabah (Sabah, 1989, p. 48, [51]) sont mis en oeuvre en fonction des applications (*cf.* partie 1.3.1.4 TYPOLOGIE DES APPLICATIONS, page 50).

Les applications qui traitent l'oral (comme la reconnaissance de la parole) commencent obligatoirement par une analyse phonologique. Quant aux applications qui traitent la langue écrite, comme les moteurs de recherche, l'analyse linguistique commence au niveau morphologique. Une analyse phonologique peut intervenir afin de proposer la tolérance aux fautes d'orthographe ou la recherche sur des textes scannés avec reconnaissance optique des caractères (OCR).

Les traitements linguistiques effectués par les moteurs de recherche seront approfondis dans la partie 1.3.3.1 OPÉRATIONS EFFECTUÉES PAR LES MOTEURS DE RECHERCHE LINGUISTIQUES (VERITY K2), page 62. Ces opérations se situent surtout aux niveaux suivants :

- Analyse morphologique
- Analyse lexicale
- Analyse syntaxique
- Analyse sémantique



Exemple d'architecture en série (d'après Sabah, 1988, P.48)

Source : Gérard Sabah, L'intelligence artificielle et le langage, vol. 2 : Processus de compréhension, Hermès, 1989 (Sabah, 1989, p. 48, [51] ; repris par Marchand, 1998, p. 17, [52])

Figure 5 – Niveaux d'analyse de la langue (TALN)

Mais, dans la pratique, ces niveaux peuvent être fortement imbriqués. On parle, par exemple, d'analyse morpho-syntaxique, ce qui englobe les niveaux morphologique et syntaxique, mais aussi lexical ; on parle aussi d'étiquetage morpho-syntaxique, défini par P. Paroubek et M. Rajman (Pierrel, 2000, p. 135, [53]) :

> *« L'étiquetage morpho-syntaxique peut être vu comme un processus en trois étapes : segmentation de la chaîne de caractères en mots (en anglais, **tokenization**), étiquetage **a priori** [ou hors contexte, n.a.] (en anglais **lexical lookup**) correspondant à la production, pour chacun des mots ainsi identifiés, de l'ensemble des étiquettes morpho-syntaxiques candidates et désambiguïsation (en anglais, **disambiguation**) correspondant à la sélection, pour chacun des mots et en fonction de son contexte, de l'étiquette morpho-syntaxique pertinente. »*

Les analyses qui interviennent dans le traitement linguistique effectué par les moteurs de recherche sont :

❑ Analyse morphologique

La morphologie est la partie de la linguistique dont l'objet d'étude est la forme que prennent les mots. La définition de l'analyse morphologique lors des traitements par la machine est donnée par C. Fuchs et B. Victorri (Fuchs, 1993, p. 83, [44]) :

> *« En traitement automatique, l'analyse morphologique consiste à segmenter un texte en **unités élémentaires** auxquelles sont attachées des connaissances dans le système : une fois cette segmentation effectuée, ce n'est plus le texte qui est manipulé, mais une liste ordonnée de telles unités (ou plusieurs listes, en cas d'ambiguïtés, [...]). Concrètement, pour le traitement d'un texte **écrit**, on part d'une chaîne de caractères typographiques [...] et on essaie de la découper de façon à ce que chaque segment corresponde à une unité répertoriée dans le système. »*

Le découpage du texte en mots sera abordé dans la partie 1.3.3.1.1 SEGMENTATION (DÉCOUPAGE, TOKENIZATION), page 64.

❑ Analyse lexicale

Le lexique recense les mots et leurs différents sens, indépendamment des relations qu'ils établissent dans le discours. Lors des traitements automatiques, l'analyse lexicale est un facteur important (E. Laporte, dans Pierrel, 2000, p. 25, [53]) :

> *« Le niveau lexical du traitement des langues naturelles correspond aux traitements centrés sur la notion de mot. »*

Les ambiguïtés lexicales peuvent être levées à l'aide des dictionnaires électroniques, comme, par exemple, ceux élaborés par le LADL[13] (Silberztein, 1993, [54]) ou ceux commercialisés par la société Memodata (Memodata, 1999, [55]). La lemmatisation est une des techniques qui permettent de regrouper sous la forme d'entrée dans le dictionnaire les formes dérivées (dérivation flexionnelle). Elle donne de bons résultats pour le traitement du français ; une autre technique est la dérivation inverse, plus adaptée à l'anglais (*cf.* partie 1.3.3.1.2 LEMMATISATION / STEMMING, page 66).

❑ Analyse syntaxique

La syntaxe étudie les relations entre les mots dans le discours (en contexte). Lors des traitements automatiques, l'analyse syntaxique peut aller jusqu'à construire des graphes représentant les relations grammaticales. La définition donnée par C. Fuchs et B. Victorri (Fuchs, 1993, p. 105, [44]) est la suivante :

[13] Les dictionnaires électroniques élaborés par l'équipe de Maurice Gross au LADL (Laboratoire d'Automatique Documentaire et Linguistique) sont : le DELAS (Dictionnaire Électronique du LADL pour les mots Simples), qui compte actuellement 90.000 entrées, et le DELAC (Dictionnaire Électronique du LADL pour les mots Composés), lexique contenant 100.000 mots composés. Le DELAS est accompagné de la transcription phonétique des mots (DELAP) et d'un système de flexion automatique qui permet de constituer le dictionnaire complet des formes fléchies (DELAP). Commencés dès les années 1970, ils font partie du système INTEX (Silberztein, 1993, [54]).

« En traitement automatique, l'analyse syntaxique consiste à associer à la chaîne découpée en unités, une représentation des groupements structurels entre ces unités ainsi que des relations fonctionnelles qui unissent les groupes d'unités. »

❑ Analyse sémantique

La sémantique a pour but l'étude du sens du discours. En traitement automatique, l'unité maximale est la phrase, voire, selon les systèmes, uniquement des unités plus petites, comme, par exemple, les groupes nominaux. C. Fuchs et B. Victorri (Fuchs, 1993, p. 139, [44]) fournissent cette définition :

*« En traitement automatique, l'analyse sémantique consiste à associer à une séquence de marqueurs linguistiques (de longueur variable) une « représentation interne » censée consigner le **sens** de cette séquence. »*

L'analyse sémantique s'appuie sur les résultats des analyses précédentes, comme le font remarquer les mêmes auteurs (Fuchs, 1993, p. 140 , [44]) :

*« Les séquences linguistiques (phrases) dont l'analyseur sémantique doit décrire le sens se composent d'un certain nombre de **mots** identifiés par l'analyse morphologique, et regroupés en **structures** par l'analyse syntaxique. Ces mots et ces structures constituent autant d'**indices** pour le calcul du sens : on pourrait dire, en schématisant beaucoup, que le sens résulte de la double donnée du sens des mots et du sens des relations entre les mots ; autrement dit encore, que la sémantique se dédouble en une sémantique **lexicale** et une sémantique **grammaticale**. »*

⇒ Sémantique lexicale

Prend pour objet les « mots pleins » : noms, verbes, adjectifs. C. Fuchs et B. Victorri (Fuchs, 1993, p. 140 , [44]) la définissent ainsi :

« La sémantique lexicale est très souvent assimilée à la sémantique des « mots pleins », comme on dit, c'est-à-dire des mots qui relèvent de grandes catégories comme le substantif, le verbe et l'adjectif (par opposition aux « mots vides », mots outils grammaticaux ou mots fonctionnels comme les prépositions, les articles, etc., que l'on a plutôt tendance à traiter comme relevant de la sémantique grammaticale). »

⇒ Sémantique grammaticale

Étudie les relations entre les mots. Certains formalismes sont issus de la logique (*cf. infra*, partie 1.3.1.2.3 LA LOGIQUE, page 45) (Fuchs, 1993, p. 144 , [44]) :

« [...] la compréhension des relations entre les mots est tout aussi importante, du point de vue sémantique, que la compréhension des mots eux-mêmes. »

Les techniques de catégorisation / classification automatique de textes (*cf.* partie 1.3.3.1.7 CATÉGORISATION / CLASSIFICATION AUTOMATIQUE (AUTOMATIC CLASSIFICATION / CATEGORIZATION), page 71) sont fondées sur des analyses sémantiques (Kayser, 2001, pp. 397-424, [56]).

Le filtrage sémantique a pour but d'obtenir des résumés automatiques plus performants (Minel, 2002, [57] ; Kayser, 2001, pp. 369-395 [56] ; Chaudiron, 2001, [58] ; Pierrel, 2000, pp. 253-270, [53], chapitre écrit par J-P. Desclès et J-L. Minel) (*cf.* partie 1.3.3.1.8 RÉSUMÉ AUTOMATIQUE (AUTOMATIC SUMMARIZATION), page 73).

1.3.1.2.2 L'INFORMATIQUE

Dans le traitement automatique du langage, le développement des outils logiciels constitue sa préoccupation la plus basique. Les programmes peuvent être écrits dans des langages de programmation classiques ou dans des langages spécialisés, comme LISP (langage de traitement de listes), PROLOG (basé sur un moteur d'inférences). Ces langages, dits « de cinquième génération », sont spécifiques à l'élaboration de programmes en intelligence artificielle.

1.3.1.2.3 LA LOGIQUE

La nécessité de traiter du sens fait appel à la logique comme mécanisme déductif. Deux branches de la logique formelle[14] sont concernées : la logique des propositions et la logique des prédicats.

La logique des propositions établit des valeurs de vérité (soit vrai, soit faux) pour une proposition donnée. Les connecteurs utilisés, représentés par des signes, sont ET ($\wedge$, $\cap$), OU ($\vee$, $\cup$), NON ($\sim$, $\neg$), IMPLIQUE ($\Rightarrow$, $\supset$), ÉQUIVALENT.

Pour traiter des propositions plus complexes, le calcul des prédicats utilise, en plus, des variables (qui, en fonction des valeurs qu'elles prennent, influencent la valeur de vérité d'une proposition) et des quantificateurs : « Pour tout » ou « Quel que soit » (noté $\forall$) et « Il existe » (noté $\exists$). Ils permettent de préciser la portée de l'assertion, à savoir si elle s'applique à tous les éléments ou seulement à certains.

La logique floue (basé sur la théorie des ensembles flous) intervient aussi dans des calculs plus complexes liés à la pertinence.

1.3.1.2.4 LES MATHÉMATIQUES ET LA STATISTIQUE

Dans le traitement automatique des langues, les méthodes mathématiques et statistiques interviennent dans les méthodes de calcul des cooccurrences des mots dans un texte.

Après une première période où le langage a été traité surtout par des méthodes statistiques et mathématiques (approches statistique et procédurale), actuellement, l'approche linguistique semble s'être imposé (Chaumier, 2003, p. 15, [24]) et même les moteurs de recherche à dominante statistique intègrent des traitements linguistiques.

[14] Basée sur l'algèbre de Boole. Le mathématicien anglais George Boole (1815-1864) est considéré comme le père fondateur de la logique moderne. (*cf.* « opérateurs booléens » utilisés en recherche d'information). Ses travaux serviront de base à Charles Sanders Pierce (1839-1914), Gottlob Frege (1848-1925), Bertrand Russel (1872-1970), Alan Turing (1912-1954) et Claude Shannon (né en 1916, auteur de la théorie de l'information).

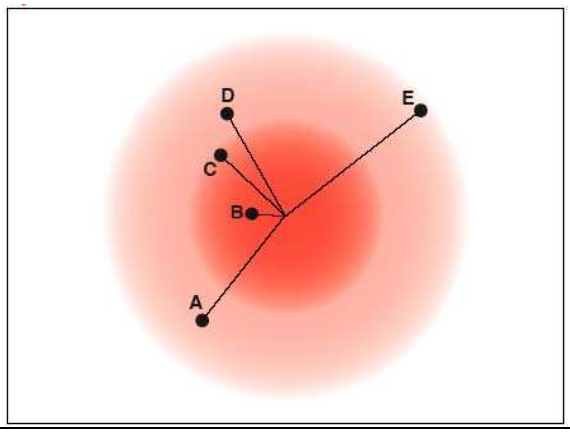
En ce qui concerne les moteurs de recherche actuels, des méthodes statistiques et mathématiques (basées sur différents algorithmes) sont utilisées pour calculer la pertinence des documents retournés par le système lors d'une interrogation, exprimée, le plus souvent, en pourcentages.

Fondée sur une vision statistique de la pertinence, la mesure de la ressemblance ou de la distance entre la requête et les documents retournés en réponse est appelée aussi coefficient de similarité.

Les modèles les plus couramment utilisés pour le calcul de la pertinence sont d'ordre booléen pondéré, vectoriel, probabiliste, et, plus récemment, basés sur les réseaux neuronaux (Lefèvre, 2000, pp. 166-176, [4] ; Pierrel, 2000, pp. 242-244, [53], chapitre rédigé par C. Fluhr) :

- **Modèle booléen pondéré**

Dans le modèle booléen classique, il n'y a pas de pondération. Les documents sont considérés soit comme complètement pertinents (s'ils contiennent tous les mots de la requête) et retournés en réponse, soit non-pertinents et, donc, ne sont pas retournés. Le modèle pondéré s'appuie sur la logique floue (voir ci-dessous Figure 6 - Ensembles flous). Pour pondérer les termes, on attribue préalablement des poids aux mots du corpus. Les termes de la requête peuvent être pondérés aussi. La fréquence du mot dans le document n'est pas prise en compte. Les réponses sont regroupées en classes (à l'intérieur desquelles tous les documents ont la même pertinence) et les résultats sont présentés par ordre décroissant de pertinence.



Source : Moteurs de recherche : où est la technologie ? (White Paper), Influo Software, 2002 (Influo Software, 2002, p. 22, [70])

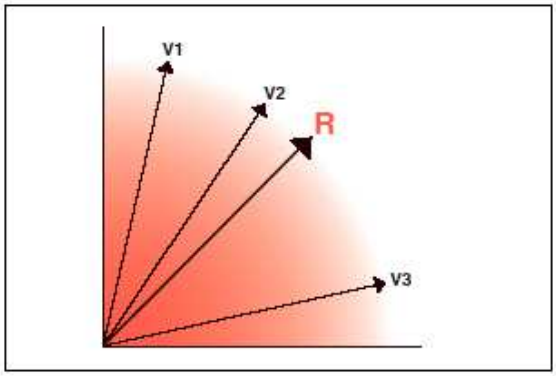
Figure 6 - Ensembles flous

❑ Modèle vectoriel

Ce modèle, créé par G. Salton, consiste à représenter le document par un vecteur d'un espace vectoriel qui comporte autant de dimensions que le nombre de mots significatifs dans le corpus (les mots vides sont exclus). Plus un terme est fréquent dans un document, plus il pèse lourd. Si un terme est peu fréquent dans le corpus, son poids sera plus important (TFIDF – Term Frequency, Inverse Document Frequency) (voir Pierrel, 2000, p. 243, [53] ; Lefèvre, 2000, p. 169, [4]). Dans les termes de P. Bellot et M. El-Bèze (Kayser, 2001, p 403, [56]) :

> «*Autrement dit, pour un document donné, un terme est important s'il apparaît souvent dans ce document et que peu de documents le contiennent.* »

La requête est représentée elle aussi comme un vecteur. Les documents dont les vecteurs sont les plus proches de celui de la requête sont considérés comme les plus pertinents. La valeur la plus couramment utilisée afin de mesurer la similarité est le cosinus de l'angle formé par les deux vecteurs (voir Figure 7 - Comparaison vectorielle des documents, page 47).



Source : Moteurs de recherche : où est la technologie ? (White Paper), Influo Software, 2002 (Influo Software, 2002, p. 25, [70])

Figure 7 - Comparaison vectorielle des documents

❑ Latent Semantic Indexing

Le modèle Latent Semantic Indexing est une variante du modèle vectoriel standard. Pour diminuer l'écart entre le vecteur de la requête et ceux des documents, en plus des mots contenus dans les documents sont ajoutées les relations sémantiques[15] implicites obtenues par la cooccurrence des termes. Pour réduire la distance entre le vecteur de la requête et ceux des documents, la matrice construite ne conserve que les *n* premiers termes (entre 100 et 300 en général, voir Jacquemin, 2000, p. 555, [59]). La recherche est effectuée à l'aide de la mesure de similarité du cosinus.

❑ Distance dans les espaces conceptuels

Le modèle vectoriel est appliqué à des espaces constitués de concepts — et non de termes, comme dans le modèle vectoriel standard.

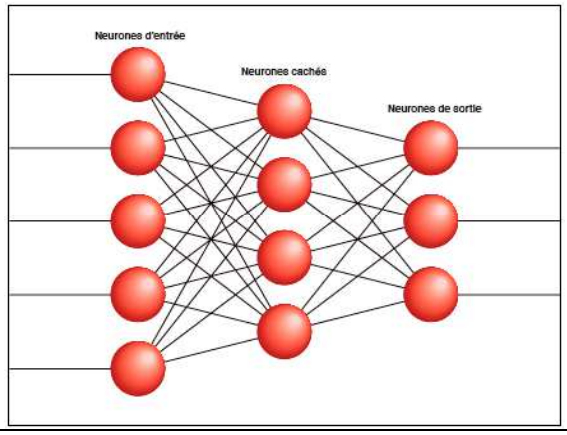
[15] Philippe Lefèvre fait remarquer que le qualificatif « sémantique » appliqué à cette méthode est contestable (Lefèvre, 2000, p. 175, [4]).

❑ Modèle probabiliste

Il s'agit d'estimer la probabilité qu'a le document d'être pertinent en réponse à la requête. Le modèle est basé sur les théorèmes de Bayes[16], concernant les probabilités conditionnelles et la probabilité des causes. La probabilité que le document soit pertinent en réponse à une requête est déduite des probabilités attribuées aux termes qui composent la requête. Donc, après des mesures effectuées sur tous les termes du corpus, on peut estimer que si un terme est présent à la fois dans la requête et dans le document, il y a une certaine probabilité que ce document soit pertinent.

❑ Réseaux neuronaux

Ils sont constitués de neurones formels, ou noeuds, généralement organisés en couches (minimum trois : requêtes-termes-documents). La requête active le réseau, ce qui entraîne la propagation des pondérations aboutissant au calcul d'un poids pour chaque document, utilisable comme coefficient de similarité (voir ci-dessous Figure 8 - Principe du réseau de neurones).



Source : Moteurs de recherche : où est la technologie ? (White Paper), Influo Software, 2002 (Influo Software, 2002, p. 27, [70])

Figure 8 - Principe du réseau de neurones

❑ Réseaux d'inférence bayésiens

Des réseaux d'inférence en plusieurs couches (documents-termes-requêtes) reçoivent des poids calculés selon les probabilités bayesiennes.

[16] Thomas Bayes, révérend et mathématicien anglais (XVIIIe siècle), a travaillé sur le calcul des probabilités. Ses théorèmes portent sur les probabilités conditionnelles et la probabilité des causes.

1.3.1.2.5 LES SCIENCES COGNITIVES

En sciences cognitives, tout système intelligent, qu'il soit naturel (humain) ou artificiel, est considéré comme une machine à traiter l'information. La modélisation et l'aspect computationnel sont considérés comme extrêmement importants. Plusieurs disciplines participent au développement des sciences cognitives, parmi lesquelles :

- Psychologie cognitive

Il s'agit, par exemple, d'expérimentations sur le comportement cognitif des sujets.

- Linguistique et psycholinguistique cognitive

La modélisation des actes d'apprentissage et d'usage de la langue s'appuie sur des expériences de laboratoire sur des groupes de sujets effectuant des tâches cognitives.

- Neurobiologie, neurosciences

La modélisation des activités humaines guidées par le cerveau, comme, par exemple, la vision lors de la reconnaissance des visages, fait appel à l'observation de sujets, sains ou atteints de lésions dans différentes zones du cerveau.

- Mathématiques et intelligence artificielle (IA)

Interviennent dans la transposition sur la machine des modélisations en conformité avec l'observation expérimentale du comportement humain.

Les implications des sciences cognitives sur le traitement du langage naturel, notamment en sémantique, et sur le traitement de l'information dans des systèmes complexes ne sont pas à négliger, notamment dans la modélisation des tâches cognitives.

1.3.1.3 OBJECTIFS DU TALN

Les finalités du traitement automatique des langues, exposées par Jean-Marie Pierrel (Pierrel, 2000, p. 18, [53]), sont de différents niveaux :

- la mise en place d'applications concrètes
- permettre de confronter la linguistique, discipline longtemps descriptive, aux exigences opératoires des traitements informatiques
- la langue, sa structure et son usage est un objet de modélisation des plus intéressants pour l'informatique linguistique (domaine privilégié de l'informatique et de l'intelligence artificielle)

1.3.1.4 TYPOLOGIE DES APPLICATIONS

Les applications du TALN (NLP) sont multiples (Chowdhury, 2003, p. 51, [47]) :

> « *Applications of NLP include a number of fields of study, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross-language information retrieval (CLIR), speech recognition, artificial intellignence, and expert systems.* »

Parmi les applications développées à partir des travaux de TALN, on peut distinguer celles qui traitent la langue parlée (l'oral) et celles qui traitent la langue écrite.

Les applications qui traitent l'oral :
- reconnaissance de la parole
- synthèse vocale

Les applications qui traitent l'écrit :
- correction orthographique
- traduction automatique (ou assistée par ordinateur)
- indexation automatique (en vue de faciliter l'accès à l'information)
- résumé automatique de textes
- reconnaissance optique de caractères (OCR – Optical Character Recognition)

L'articulation des deux permet de développer :
- le dialogue homme-machine

Le terme « industries de la langue » est utilisé pour désigner les travaux visant la fabrication de produits opérationnels (logiciels, progiciels), dans le but de les commercialiser. Ce sont des outils qui traitent des données textuelles et qui s'appuient sur les recherches en TALN appliquées à des domaines restreints. On peut ainsi citer : les correcteurs orthographiques (faisant ou non partie des traitements de texte), les logiciels d'aide à la traduction, les moteurs de recherche. On parle aussi d'ingénierie de la langue ou d'ingénierie des langues.

1.3.1.5 MATURITÉ TECHNOLOGIQUE

Dans l'introduction d'un numéro relativement récent de la revue « Traitement automatique des langues » consacré au traitement automatique des langues appliqué à la recherche d'information, Christian Jacquemin (Jacquemin, 2000, p. 328, [59]), affirme :

> *« La discipline a désormais atteint une maturité technologique suffisante pour traiter finement l'information contenue dans les masses importantes de données textuelles. »*

Cette maturité se traduit, en termes de marché, par une présence renforcée des outils commercialisés dans les entreprises. Parmi les applications qui font appel aux techniques développées par la recherche en TALN, les moteurs de recherche tiennent une place de plus en plus importante. Leur succès sur le web n'est plus à démontrer et les entreprises (surtout les grands comptes) se dotent d'intranets où le besoin d'un tel outil se fait très vite sentir.

1.3.1.6 LIMITES

Quant à la compréhension du langage humain par les ordinateurs, les réserves émises par Philippe Lefèvre (Lefèvre, 2000, p. 44, [4]) il y a quelques années semblent toujours d'actualité :

> *« Il ne faut donc pas s'attendre à ce qu'un programme de traitement du langage naturel puisse extraire d'un texte des concepts qui n'auront pas été prévus par les concepteurs du système, ou qu'il puisse construire des modèles nouveaux de représentation de l'information. Tout ce qu'il est possible d'obtenir pour l'instant, c'est la mise en en évidence d'associations nouvelles de concepts déjà introduits dans le système, ou l'actualisation de règles de grammaire déjà entrées, de modèles prédéfinis »*

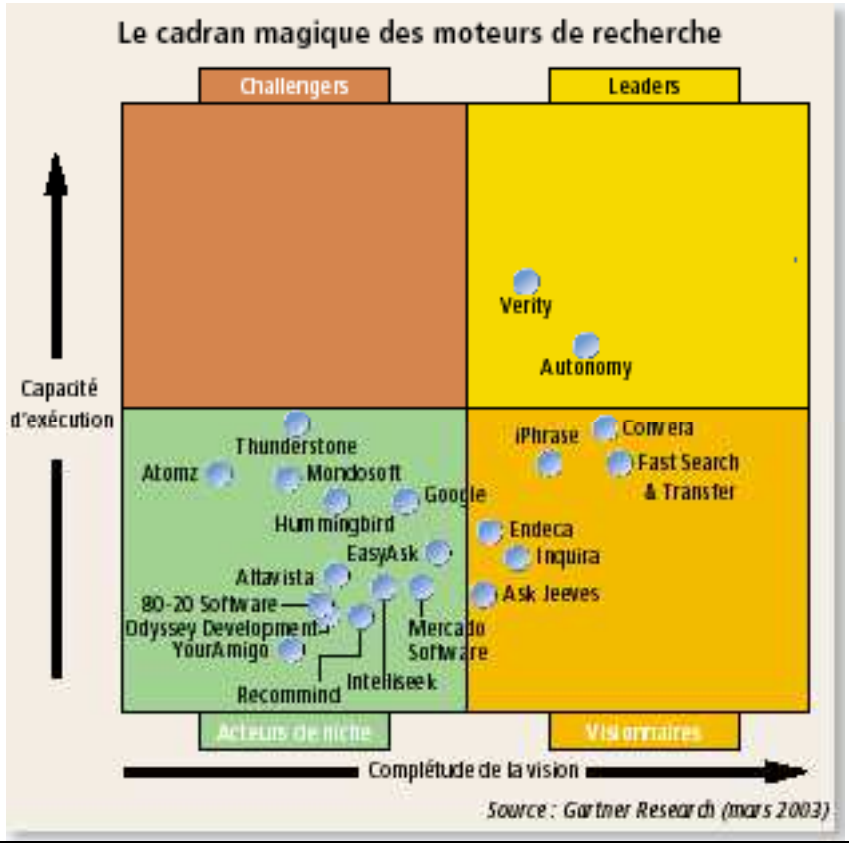
La compréhension du texte par la machine serait ainsi la mise en correspondance d'un modèle prévu avec des textes nouveaux : plus le modèle est sophistiqué, meilleurs sont les résultats.

D'autre part, (Lefèvre, 2000, p. 44, [4]), le développement des réseaux neuromimétiques, des bio-mathématiques, des algorithmes génétiques, ouvrent des perspectives de développement concernant les capacités d'apprentissage de la machine.

1.3.2 POSITIONNEMENT DES ACTEURS SUR LE MARCHÉ ET TYPOLOGIE DES PRODUITS

1.3.2.1 MARCHÉ DES MOTEURS DE RECHERCHE

Un article paru au mois de septembre dans 01 Informatique (Roberget, 2003, p. 7, [60]), reprenant les chiffres publiés par Gartner Research en mars 2003, présente la cartographie des acteurs présents sur le marché des moteurs de recherche.



Source : Olivier Roberget, Le marché impose son modèle économique à la recherche sur internet, 01 Informatique, no. 1735, 5 septembre 2003 (Roberget, 2003, p. 7, [60])

Figure 9 - Positionnement des éditeurs de moteurs de recherche sur le marché

Malgré un marché relativement atone dû à la crise économique, le chiffre d'affaires de certains a toutefois augmenté. D'autre part, les rachats se sont poursuivis (cinq sociétés ont été reprises entre novembre 2002 et juillet 2003), autant dans le domaine des activités web que dans le domaine de la recherche en entreprise, où la problématique est différente.

1.3.2.2 TYPOLOGIE DES MOTEURS DE RECHERCHE

Selon le choix des technologies (linguistiques ou statistiques) mises en oeuvre de façon prépondérante par les concepteurs du logiciel, une typologie des moteurs de recherche peut être établie (Chaumier, 2003a, [61] ; Crochet Damais, 2003, [62] ; Chilotti, 2003, [63] ; Dalbin, 2002, [64] ; Girardeau, 2002, [65] ; Crochet Damais, 2002, [66] ; Lefèvre, 2000, [4] ; Leloup, 1998, [43]). La plupart des acteurs présents en ce moment sur le marché français peuvent être classés selon ces critères. D'autres solutions logicielles sont citées dans différentes sources, mais leur présence dans les entreprises est pour l'instant marginale.

1.3.2.2.1 MOTEURS DE RECHERCHE STATISTIQUES

Tous les moteurs de recherche utilisent des méthodes statistiques. Mais les produits rangés dans cette catégorie accordent la place prépondérante à l'analyse statistique par rapport à l'analyse linguistique.

Les moteurs de recherche statistiques mettent en oeuvre, de façon prioritaire, des méthodes de calcul statistiques basées sur l'occurrence et la co-occurrence des mots dans le texte, comparés à sa fréquence dans le corpus. Des algorithmes différents peuvent servir pour la recherche des documents et pour générer une taxonomie. Ils ne contiennent généralement pas de dictionnaires de langues et la plupart se disent indépendants de la langue des documents.

Ce type d'outils est particulièrement adapté à des grands corpus hétérogènes, comme, par exemple, les bases de presse, ce qui permet aux algorithmes statistiques d'établir les relations entre les chaînes de caractères co-occurrentes dans plusieurs documents et de leur attacher des poids en fonction de leur fréquence d'apparition dans la base.

Ils sont adaptés aussi aux sites Internet des grandes entreprises, lorsque les services estiment qu'il est difficile de prévoir le type de questions qu'un internaute pourrait poser, ce qui exclut la construction d'un thésaurus ou d'une taxonomie destinés à orienter les requêtes.

Ce type de solutions logicielles sont moins adaptées aux corpus métier, car ceux-ci contiennent des documents dans lesquels plusieurs termes peuvent être présents sans pour autant rendre le document pertinent pour une requête sur l'un d'entre eux.

Ils ne sont pas du tout adaptés aux petits corpus, car les algorithmes statistiques ont besoin d'une masse importante de documents pour faire des calculs basés sur la fréquence d'apparition des chaînes de caractères et établir des relations pour donner des résultats pertinents.

Les principaux moteurs de recherche statistiques sont :

- **Autonomy (Autonomy)**

Parmi ce type d'outils, Autonomy (IDOL Server, Autonomy Server, Portal-in-a-Box) de la société Autonomy est le mieux placé sur le marché des moteurs de recherche (voir Figure 9 - Positionnement des éditeurs de moteurs de recherche sur le marché, page 52) : c'est le 2e acteur du marché avec un chiffre d'affaires de 51 millions de dollars pour l'exercice fiscal clos fin mars 2003[17] (malgré une baisse de 3%) (Roberget, 2003, p. 7, [60]). Il est basé sur des algorithmes bayesiens et sur la théorie de Shannon, selon laquelle plus un terme est rare dans le corpus, plus son poids est important. Son fonctionnement est indépendant des langues et il génère un réseau de concepts sans intervention humaine. La société a 575 clients dans le monde[18], parmi lesquels : General Motors, Lucent Technologies, AT&T, Ericsson, Novartis, Sun Microsystems, BAE Systems, Bell South, Unisys, BBC et Reuters.

- **Exalead (Exalead)**

Sur le marché français, on peut citer Exalead, de la société Exalead, essentiellement statistique, mais qui intègre des traitements d'ordre linguistique (« lemmatisation statistique ») et génère un classement du lot-résultat à la volée. Cette solution logicielle a été choisie notamment par AOL France et la Société Générale (pour son site Internet).

- **MatchPoint (TripleHop)**

MatchPoint (société TripleHop) utilise des algorithmes statistiques (algorithme SVM propriétaire) et l'apprentissage de la machine, mais intègre une recherche par concepts et une taxonomie prédéfinie avant l'installation de l'application (JC Decaux, AOL Time Warner). (TripleHop Technologies, 2002, [67])

1.3.2.2.2 MOTEURS DE RECHERCHE LINGUISTIQUES (ET SÉMANTIQUES)

Les outils rangés dans cette catégorie mettent l'accent sur les traitements linguistiques. Parmi les traitements d'ordre linguistique mis en oeuvre, ceux basés sur la morphologie et la syntaxe sont présents dans tous les systèmes actuels, mais la sémantique peut jouer un rôle plus ou moins important. Les méthodes statistiques sont utilisées ici pour accélérer le traitement des textes ou participent aux calculs de pertinence lors de la présentation des résultats de recherche, mais la place qu'elles tiennent dans ce type d'outils n'est pas prépondérante.

Les moteurs linguistiques sont particulièrement adaptés aux corpus métier (qu'ils soient petits ou grands), contenant des termes précis et ciblés. Une gestion rigoureuse des thésaurus, réseaux sémantiques, dictionnaires métier, référentiels d'entreprise, taxonomies et concepts est nécessaire, qu'ils soient mis en place dans l'application par l'éditeur (Sinequa), par l'utilisateur (Verity), ou vendus en option (RetrievalWare).

Les principaux moteurs de recherche linguistiques / sémantiques sont :

[17] Tous les chiffres concernant les résultats comptables des éditeurs sont extraits de l'article d'Olivier Roberget paru dans le journal 01 Informatique du 5 septembre 2003 [60]

[18] Chiffres de l'éditeur (2001).

❑ K2 Enterprise (Verity)

Le moteur de recherche leader du marché, K2 Enterprise de la société américaine Verity, se situe dans cette catégorie. Le chiffre d'affaires de l'entreprise a augmenté de 9 % sur l'exercice clos fin mars 2003 et se situe à hauteur de 102 millions de dollars (voir Figure 9 - Positionnement des éditeurs de moteurs de recherche sur le marché, page 52). Ses principaux atouts : 250 formats pris en charge, 26 langues traitées (dictionnaires syntaxiques couplés à des règles de grammaire), la rapidité des réponses, la volumétrie. C'est, à la base, un moteur d'indexation full-text, possédant plus de 35 opérateurs, enrichi d'une indexation par concepts et taxonomie. La société Verity a plus de 3500 clients dans le monde dont plus de 500 en France[19], parmi lesquels on peut citer : AFP, Airbus, Air France, Alcatel, Aventis, Castorama, CNDP, EDF, Dassault Aviation, Éditions Dalloz, Éditions Lamy, France Telecom, Hachette Filipacchi, Lapeyre, La Poste, Le Ministère de l'Économie, Inist/CNRS, Pechiney, Renault, Société Générale, Thalès.
Un deuxième produit, destiné aux PME/PMI, issu du rachat des activités entreprise d'Inktomi en novembre 2002, est commercialisé sous le nom de **Verity Ultraseek**. (Le moteur Inktomi Search traitait 35 langues et comptait environ 2500 grands comptes parmi ses clients.)

❑ RetriewalWare (Convera)

RetrievalWare, le moteur de recherche fourni par la société Convera, troisième acteur du marché (voir Figure 9 - Positionnement des éditeurs de moteurs de recherche sur le marché, page 52) (société appelée auparavant Excalibur ; chiffre d'affaires : 23,6 milliards de dollars, en baisse de 31% sur l'exercice clos fin janvier 2003), supporte plus de 200 formats et a la particularité de pouvoir traiter les formats multimédia et le multilinguisme (6 langues : anglais, espagnol, français, allemand, italien, russe, japonais, chinois, arabe). L'indexation se fait sur les données binaires, mais elle peut être enrichie par un thésaurus existant ou par des des dictionnaires métier développées par l'éditeur (« cartouches linguistiques »). La société a presque 1.000 clients et partenaires dans le monde[20], parmi lesquels : Air France, AT&T, AUDI, Boeing, Czech Telecom, Deutsche Post, Deutsche Telekom, Eutelsat, Ford Motor Company, Framatome, General Electric, Nortel Networks, PSA Peugeot Citroën, Telefonica, Unilever, Unisys, United Airlines, Xerox. (voir aussi Villacampa, 1999, [68])

❑ Arisem KM Server (Arisem)

Un des principaux acteurs du marché français est la société Arisem Group, dont le produit actuel s'appelle Arisem KM Server. Le produit antérieur, OpenPortal4U, gérait 200 formats et 9 langues (dont 5 en cross-language), mais ne pouvait pas effectuer de recherches sur Internet et pouvait être installé uniquement sur des serveurs Windows NT. Ses traitements linguistiques (morpho-syntaxiques et sémantiques) sont basés sur 20.000 concepts (thésaurus) et 400.000 règles et il est possible d'enrichir ce dictionnaire. L'entreprise compte environ 50 clients au niveau mondial, dont environ 45 en France. On peut citer : Alcatel, CNES (Centre Nationale d'Études Spatiales), EADS, EDF, GDF, France Telecom, Ministère de l'Économie des Finances et de l'Industrie, Pernod-Ricard, Radio France, Saint-Gobain, Thalès, TotalFinaElf, Unilog, Usine Nouvelle.

[19] Chiffres de l'éditeur.

[20] Chiffres de l'éditeur.

❑ Intuition (Sinequa)

Intuition (société Sinequa, anciennement Cora) reconnaît plus de 200 formats et supporte 13 langues, parmi lesquelles : français, anglais, allemand, espagnol, italien, néerlandais, japonais, chinois traditionnel, thaïlandais (interlinguisme pour les langues européennes). Ses traitements linguistiques (syntaxiques et sémantiques) sont fondés sur un lexique de 250.000 formes par dictionnaire et 100 règles de grammaire. Le moteur de recherche Intuition est installé chez environ 40 clients en France[21], parmi lesquels : AlloCiné, Benchmark group, Cedocar, Commission des Opérations de Bourse (COB), Diva Press, EDF, Encyclopaedia Universalis, Journal du Net, La Redoute, Le Monde, Leroy Merlin, Les 3 suisses, Ouest France, Reporters Sans Frontières, Saint-Gobain, SNCF.

❑ Spirit (T-Gid)

Spirit, commercialisé par la société T-Gid, traite 250 formats (qu'il transforme en HTML) et 5 langues : français, anglais, allemand, néerlandais, russe. Son interface est personnalisable, mais l'indexation semble lente. Ses ressources linguistiques comptent environ 120 règles de grammaire, un lexique de 500.000 formes pour le français et des dictionnaires sémantiques et syntaxiques pour les autres langues traitées. Parmi ses 2.300 clients[22] : AGF, Banque de France, CEA, CEDOCAR, Cetelem, Cogema, Cour des Comptes, École des Mines de Paris, EDF, Framatome, France Telecom, INSEE, Ministère de la Défense, Ministère de l'économie, des finances et de l'industrie, MMA, PPR, RFO, RTL, TF1. (voir aussi Dalbin, 2000, [69])

❑ Hummingbird SearchServer (Hummingbird)

Hummingbird SearchServer de la société Hummingbird (le produit avait été développé par la société Fulcrum avant son rachat) (voir Figure 9 - Positionnement des éditeurs de moteurs de recherche sur le marché, page 52). C'est un moteur robuste et rapide, capable de traiter des volumes importants (comme les 8.000.000 de notices de la BnF). Les langues prises en charge sont : français, anglais, allemand, néerlandais, italien, espagnol, portugais, norvégien, suédois, danois, chinois, japonais et coréen ; pour certains il existe des thésaurus. Parmi ses réalisations en Franceet en Europe : le catalogue en ligne de la Bibliothèque nationale de France (BnF), CCF, Carrefour (boutique en ligne), la base documentaire de la Commission Européenne. La société propose d'autres outils, notamment de gestion de contenu, et compte parmi ses clients, au niveau mondial : Alcatel, AXA Group, Barclays Bank, Bell Canada, BMW, BNP Paribas, Chrysler Corp, Deutsche Bank, Land Rover, NASA, US Army, US Navy, Verizon.

❑ Lingway Knowledge Management Tools (Lingway)

LKMT (Lingway Knowledge Management Tools) est une suite logicielle intégrant des fonctions de gestion des connaissances développée par la société Lingway (créée en 2001 par des anciens de Lexiquest/ERLI). Articulée en plusieurs modules, la suite permet de mettre en place des fonctions de recherche en langage naturel (modules Fulty et Tacsy), catégorisation, extraction de données et génération de texte, intégrables selon les besoins exprimés. Les solutions de Lingway peuvent traiter 8 langues, dont le français, l'anglais, l'allemand et l'espagnol. Parmi ses clients :

[21] Chiffres de l'éditeur.

[22] Chiffres de l'éditeur.

EADS, INPI, INSEE, Office européen des brevets, Umanis Clinical Research, Instituto Nacional de Meteorologia.
Les deux autres solutions logicielles proposées sont **Lingway Patent Suite** (accès multilingue à des bases de brevets) et **Lingway Medical Dictionary Encoder** (assistant de recherche : effets secondaires des médicaments).

1.3.2.2.3 ASSISTANTS (OU FÉDÉRATEURS) DE REQUÊTES

Les logiciels classés dans cette catégorie ne font pas forcément la recherche par eux-mêmes : ils s'appuient sur des moteurs de recherche ou s'interfacent avec les solutions logicielles présentées plus haut via des partenariats. Ces logiciels traduisent les requêtes de l'utilisateur dans les différents langages d'interrogation propriétaires ou dans les protocoles d'interrogation des bases de données en ligne (SQL, Z39.50). Leur utilisation permet à l'entreprise d'utiliser un grand nombre de sources, internes ou externes, à travers une interface unique. Certains gèrent aussi les droits d'accès en s'appuyant sur les annuaires LDAP des entreprises. Parmi ces solutions, on peut citer :

- Albert Meaning Interpreter (AMI) de la société Albert
- Lexiquest Guide de la société SPSS (anciennement Lexiquest, ERLI)
- AskOnce (développé par le Xerox Research Center Europe)

1.3.2.2.4 QUELQUES AUTRES ACTEURS ET AUTRES APPROCHES

Quelques autres acteurs : Auraweb (société Auracom), Aperto Libro (Alogic), Co-Brain et Knowledgist (Invention Machine), Kartoo (Kartoo), Knowings KMS (Knowings), Online Miner (Temis), Pertimm (Systal Pertimm), Synomia Site Search (Synomia), Tetralogie (IRIT-Innovation), Tropes (Acetics), Imap/See-K (Trivium).

- Quelques autres approches peuvent être citées :

Influo (Influo Software) prétend mettre en oeuvre des technologies basées sur une « technologie d'inspiration neuro-biologique », qui « s'appuie sur une modélisation neuro-biologique du raisonnement humain ». (Influo Software, 2002, p. 1, [70] ; Villoing, 2003, [71])

Le moteur de recherche dit « multidimensionnel » Instranet de la société Instranet est une plateforme de gestion de contenu maketing (BNP, Crédit Lyonnais).

- Des modèles économiques différents se développent :

Des solutions qui fonctionnent en mode ASP sont apparues, comme, par exemple, Antidot Finder Suite de la société Antidot (site web Journal du Net), Atomz de la société Atomz (CBS, Macromedia), Synomia Site Search de Synomia (CFDT).

Récemment commencent à se développer des moteurs Open Source, dont voici quelques exemples : Jakarta Lucene, Java Search Engine, ASPseek, mnoGoSearch.

1.3.3 FONCTIONNEMENT DES MOTEURS DE RECHERCHE LINGUISTIQUES : L'EXEMPLE DE VERITY K2

Malgré l'extrême complexité atteinte par les moteurs de recherche, pour une utilisation optimale des systèmes commercialisés actuellement, selon Christian Fluhr[23], expert en ingénierie linguistique, il est nécessaire de faire connaître aux utilisateurs les « étapes de raisonnement » (Roumieux, 2000b, p. 24, [72]) :

> *« Pour une bonne acceptabilité des systèmes, il faut quand même que l'utilisateur ait une certaine compréhension de la manière dont le système fonctionne, qu'il puisse avoir une maîtrise de ce qui se passe. »*

Pour ce faire, si l'on revient à la définition d'un moteur de recherche comme étant un programme qui indexe des documents et stocke cet index pour le parcourir ensuite en fonction de la requête de l'utilisateur (voir définition des moteurs de recherche, page 38), la première phase sera donc l'indexation.

Pour Verity, cela s'appelle constituer une (des) collection(s) (Verity, 2001c, p. 3-2, [73]) :

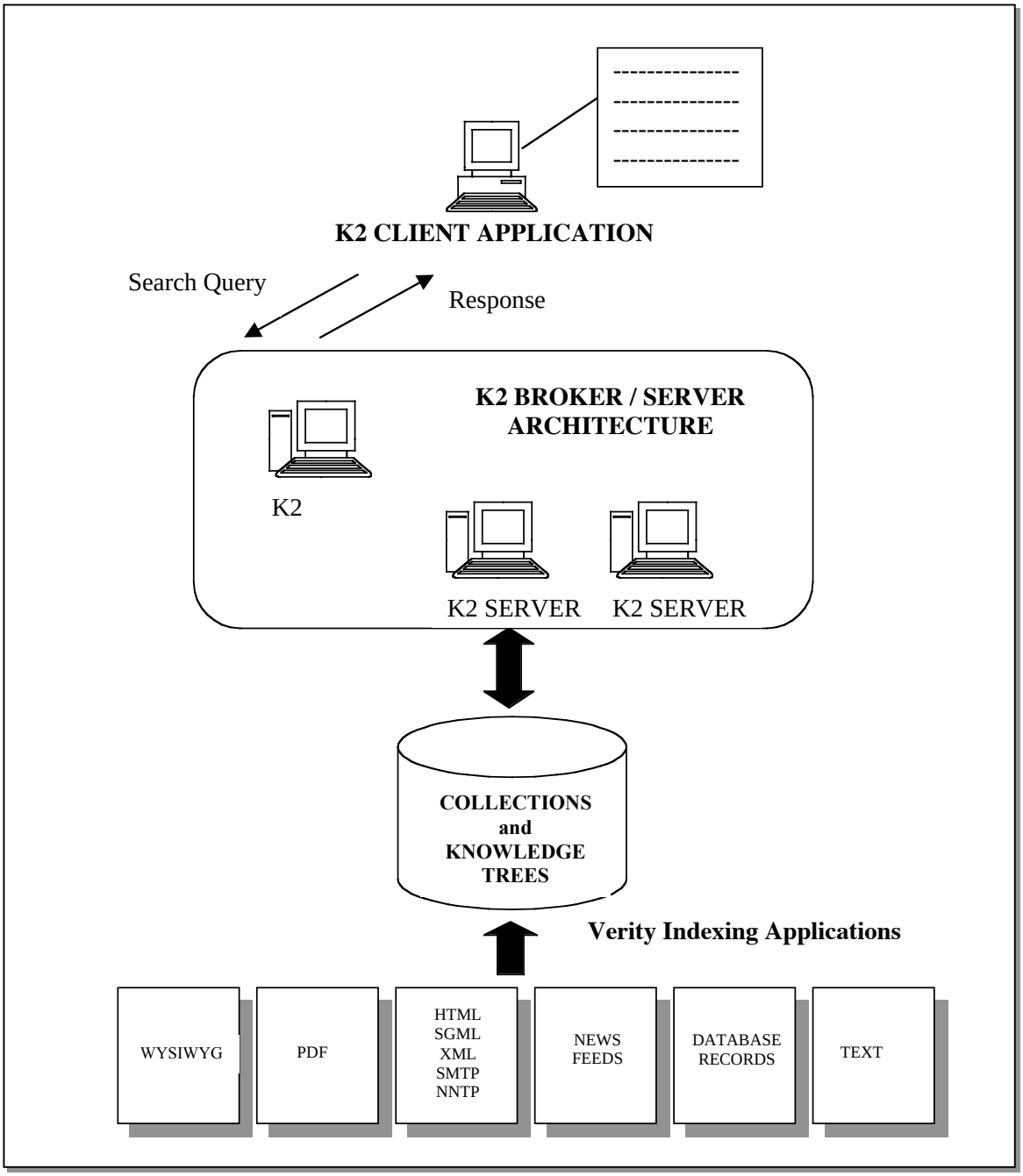
> *« A collection is indexed information about a set of documents, stored in a directory structure. Collection content and configuration support basic and advanced search, navigation, and view functions. »*

Les documents sont indexés et laissés à leur place, le moteur conserve uniquement ses fichiers index en format propriétaire, les collections, qui sont utilisées pour fournir les réponses à une requête. (voir Figure 10 - Schéma de l'architecture d'une application Verity, page 59).

Les données et les métadonnées des documents sont extraites, filtrées, traitées (du point de vue linguistique), indexées (du point de vue informatique) et stockées pour permettre ensuite la recherche (voir Figure 11 - Index et filtres (visualizeur universel KeyView, filtres PDF, HTML), page 61, et Figure 12 - Opérations effectuées par le moteur Verity (exemple d'application multilingue), page 63).

[23] Christian Fluhr est un des fondateurs de T-Gid, éditeur du moteur de recherche Spirit.

Architecture d'une application Verity



Source : Verity K2 Search Objects Guide, Verity, 2001, p. 1-2 (Verity, 2001e, [74])

Figure 10 - Schéma de l'architecture d'une application Verity

❑ Formats de fichiers pris en charge par Verity

Du point de vue informatique, un fichier doit respecter certaines règles d'écriture afin d'être compris par un ordinateur dans le cadre d'un échange entre programmes. Tous les formats de fichiers ne sont pas lisibles directement à l'écran par un humain, car ils n'ont pas été conçus dans ce but. Certains formats sont destinés à la communication entre deux machines comme, par exemple, un ordinateur et une imprimante (PostScript d'Adobe) ou au transfert de données, comme KIF (Knowledge Interchange Format).

Selon les formats de fichier, on parle de documents (ou information) structurés (XML, tables des bases de données relationnelles) ou non-structurés (Word, HTML). Dans un document structuré, les informations de même nature (exemple trivial : numéros de téléphone dans une base contacts ou noms d'auteurs dans une base bibliographique) se trouvent décrits toujours de la même façon. En termes informatiques, cette description s'appelle la syntaxe et, dans le cas d'un document structuré, elle est utilisable pour la recherche des informations.

Lorsque cette syntaxe suit des normes internationales et des standards (ex. : SGML, HTML, XML) plutôt que des formats propriétaires, la recherche sur des documents générés à l'aide de logiciels différents se trouve grandement facilitée.

Les formats propriétaires posent problème, car, pour analyser un texte, il faut d'abord pouvoir le lire, assertion valable autant pour un humain que pour une machine. Dans le cas du traitement automatique du langage par un moteur de recherche, le problème peut être résolu de manière différente, selon les éditeurs.

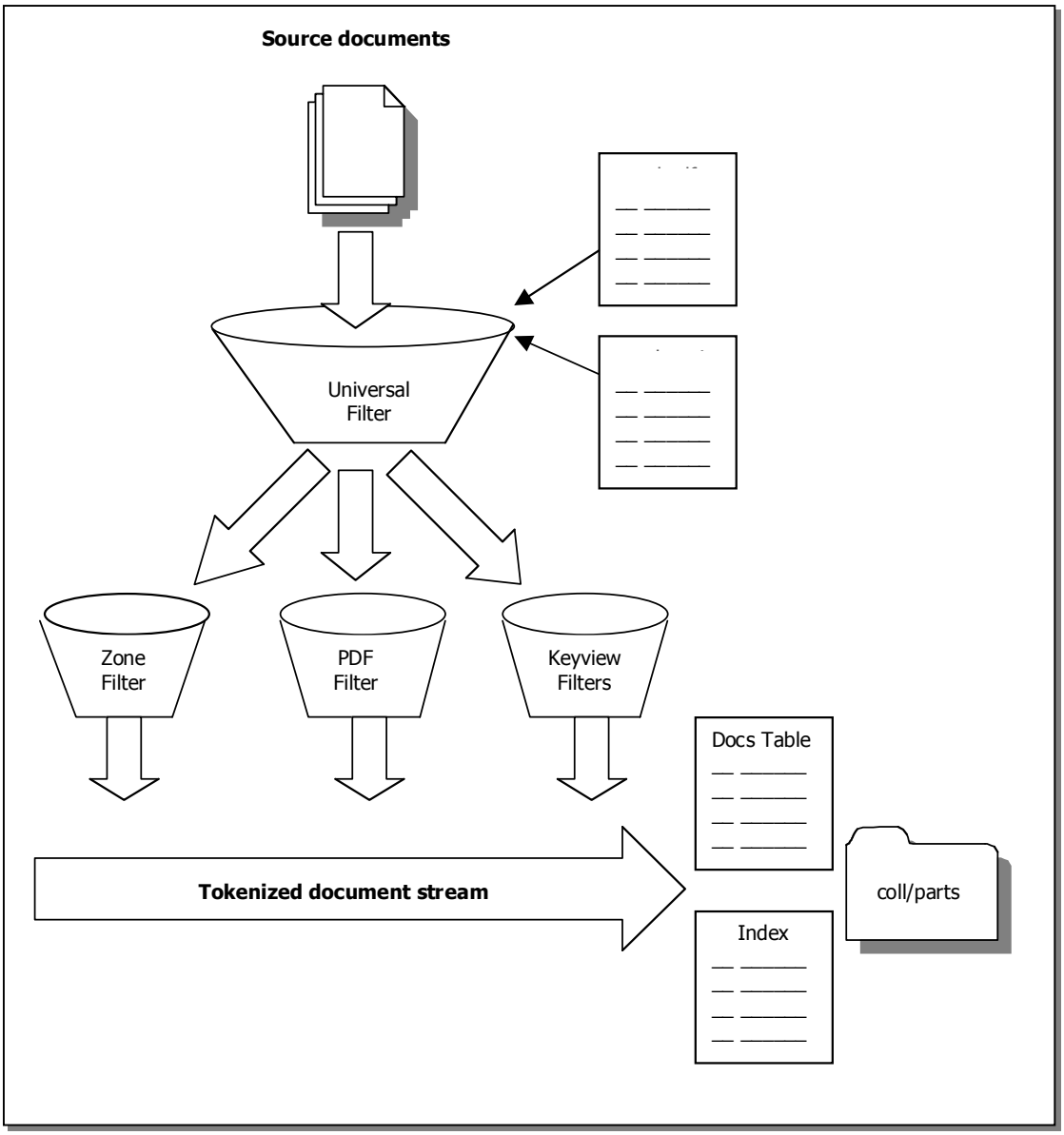
La plupart des éditeurs de moteurs de recherche font appel à un type de logiciel particulier, appelé « visualizeur universel », qui permet la lecture et l'affichage des formats générés par des logiciels qui ne sont pas installés sur la machine client.

L'éditeur du visualiseur universel KeyView a été racheté, il y a quelques années, par la société Verity. Il est actuellement partie intégrante du moteur de recherche K2 Enterprise. La visualisation de plus de 250 formats différents se fait à travers des filtres (voir Figure 11 - Index et filtres (visualizeur universel KeyView, filtres PDF, HTML), page 61).

❑ Langues traitées

Le moteur de recherche Verity K2 Enterprise peut traiter 26 langues via l'installation d'un ou plusieurs des 26 modules linguistiques (appelés « locales ») développés par des partenaires locaux de l'éditeur. Les langues traitées (en 2001) sont : allemand, anglais, arabe, bulgare, chinois simplifié, chinois traditionnel, coréen, danois, espagnol, finnois, français, grec, hébreu, hongrois, italien, japonais, néerlandais, norvégien bokmal, norvégien nynorsk, polonais, portugais, russe, suédois, tchèque, thaï, turque (Verity, 2001b, p. 5, [75]).

Les traitements linguistiques sont différents pour chaque langue, mais on peut intervenir dans le paramétrage de certaines fonctions comme, par exemple, la liste des mots vides et les caractères spéciaux.



Source : Verity K2 Enterprise Fundamentals, Verity, 2001, p. 9-5 (Verity, 2001c, [73])

Figure 11 - Index et filtres (visualizeur universel KeyView, filtres PDF, HTML)

1.3.3.1 OPÉRATIONS EFFECTUÉES PAR LES MOTEURS DE RECHERCHE LINGUISTIQUES (VERITY K2)

La description des opérations effectuées par les moteurs de recherche s'appuie sur la littérature concernant les types d'analyses linguistiques mises en oeuvre en TALN (cf. partie 1.3.1 MOTEURS DE RECHERCHE ET TRAITEMENT AUTOMATIQUE DES LANGUES « NATURELLES » (TALN), page 40) et sur la typologie de l'indexation établie par Philippe Lefèvre[24] (Lefèvre, 2000, p. 108-127, [4]).

Les exemples concernant la description des opérations effectuées Verity K2 Enterprise s'appuient sur la documentation de l'éditeur (fournie avec la version 4.0.1 du moteur et les « White Papers » disponibles sur le site), ainsi que sur l'expérience acquise en création de concepts et de taxonomie à l'aide de l'interface dédiée Intelligent Classifier (voir partie 1.3.3.2 INTELLIGENT CLASSIFIER, MODULE DE GESTION DES CONCEPTS ET DE LA TAXONOMIE, page 75).

Les principales opérations effectuées par les moteurs de recherche seront présentées en mentionnant les niveaux d'analyse linguistique qui interviennent (voir aussi partie 1.3.1.2.1 LA LINGUISTIQUE, page 41).

Les opérations qui mènent à la constitution de la collection Verity sont présentées dans le schéma page 63, Figure 12 - Opérations effectuées par le moteur Verity (exemple d'application multilingue).

Selon les niveaux d'analyse linguistique qui interviennent, ces opérations sont :

NIVEAUX D'ANALYSE MORPHOLOGIQUE ET LEXICALE
- segmentation
- lemmatisation

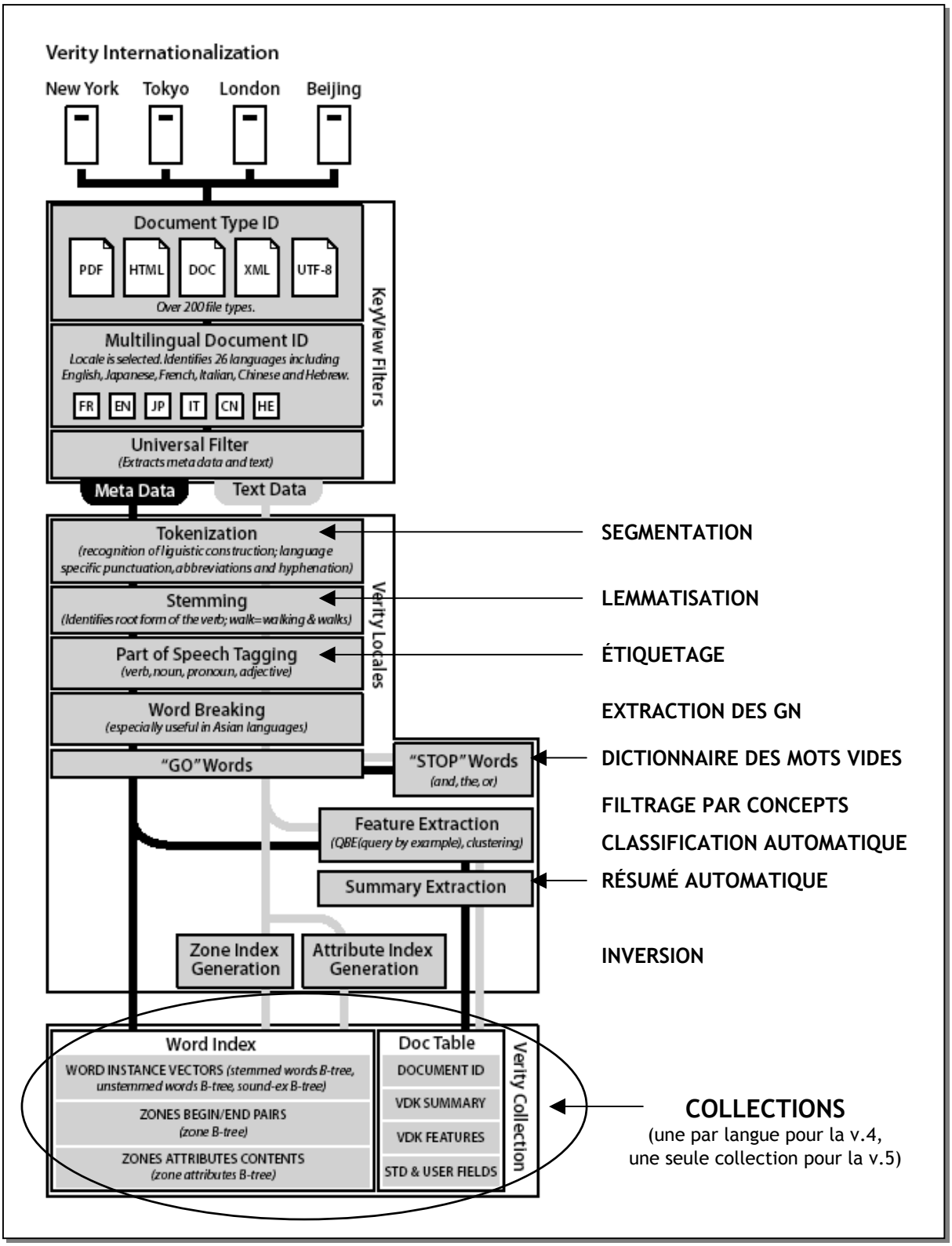
NIVEAU D'ANALYSE SYNTAXIQUE
- étiquetage
- extraction des groupes nominaux

NIVEAU D'ANALYSE SÉMANTIQUE
- filtrage par concepts
- classification automatique
- résumé automatique

(élimination des mots vides – paramétrage facultatif)
(inversion – opération informatique)

[24] Toutefois, le livre de Lefèvre est paru en 2000 et la version du moteur de recherche de Verity citée en exemple dans l'ouvrage est celle dont le nom commercial est Verity Information Server, version inférieure qui possède moins de fonctionnalités. La version décrite ici, 4.0.1, n'est pas la dernière mouture, car la version 5 est sortie en juin 2003 (et une version intermédiaire, 4.5, avait fait la transition).

OPÉRATIONS EFFECTUÉES PAR LE MOTEUR VERITY (exemple d'application multilingue)



Source : Verity Internationalization : Enabling E-business in multiple Languages : Verity White Paper, Verity, 2001, p. 5 (Verity, 2001b, [75])

Figure 12 - Opérations effectuées par le moteur Verity (exemple d'application multilingue)

1.3.3.1.1 SEGMENTATION (DÉCOUPAGE, TOKENIZATION)

La première opération nécessaire à l'indexation du fonds de documents par un moteur de recherche est le découpage de la chaîne de caractères en unités lexicales (Habert, 1997, p. 162, [46]).

> *« La segmentation du texte consiste à découper une suite de caractères en « unités » : mots simples ou unités polylexicales. »*

Appelée segmentation (ou découpage), l'opération consiste à segmenter le texte en unités élémentaires (« token » en anglais) : mots, phrases, paragraphes. L'opération s'effectue à l'aide d'un programme appelé segmenteur, qui effectue un marquage des unités (« tokenizer » en anglais) selon des règles prédéfinies.

L'opération se situe au niveau morphologique. Les tokens ainsi obtenus seront stockés dans un fichier positionnel inversé, qui peut servir par la suite à mettre en surbrillance (à l'affichage des résultats de recherche) les mots de la requête dans le document affiché (Jacquemin, 2000, p. 501, [59]).

La méthode la plus évidente est de considérer que les chaînes de caractères représentent des mots et les blancs sont des séparateurs : règle « black and white tokens » (Ambroziak, 2000, p. 1, [76]). Le problème qui se pose tout de suite est celui de la ponctuation : l'approximation suivante consiste à exclure la ponctuation et à continuer le traitement selon la méthode « black and white ».

Un autre problème est celui des mots composés. « Pomme de terre » perd son sens si l'on découpe arbitrairement dès le début la chaîne de caractères en trois unités, alors qu'il y a une seule unité de sens (Pierrel, 2000, p. 135, [53]). Le problème est appréhendé de diverses façons et la désambiguïsation peut intervenir plus tard.

Quant aux mots composés à l'aide d'un tiret (comme « bas-relief »), il est important de savoir si le tiret est considéré par le logiciel lors de l'indexation comme un séparateur ou comme un caractère.

Un autre caractère qui pose problème est l'apostrophe, qui fonctionne lui aussi tantôt comme séparateur, tantôt comme composant de mots (Habert, 1997, p. 162, [46] ; Silberztein, 1993, pp. 111-119, [54]).

Les symboles (&, /, #, $, €) posent eux aussi des problèmes d'interprétation, mais effectuer une recherche sur des chaînes comme AT&T, ou OS/2 peut être indispensable lorsqu'on gère un corpus informatique.

Dans certains systèmes, il est possible de paramétrer l'application à l'installation si l'on estime que la recherche sur les mots contenant de tels caractères sont importants pour le fonds traité. La casse des mots est à prendre en compte aussi.

Un autre problème est le découpage en phrases et en paragraphes, qui s'avère délicat, car la ponctuation offre des indices peu fiables, or, ce découpage est extrêmement important pour l'analyse syntaxique et le calcul des co-occurrences.

❑ Verity et les tokens

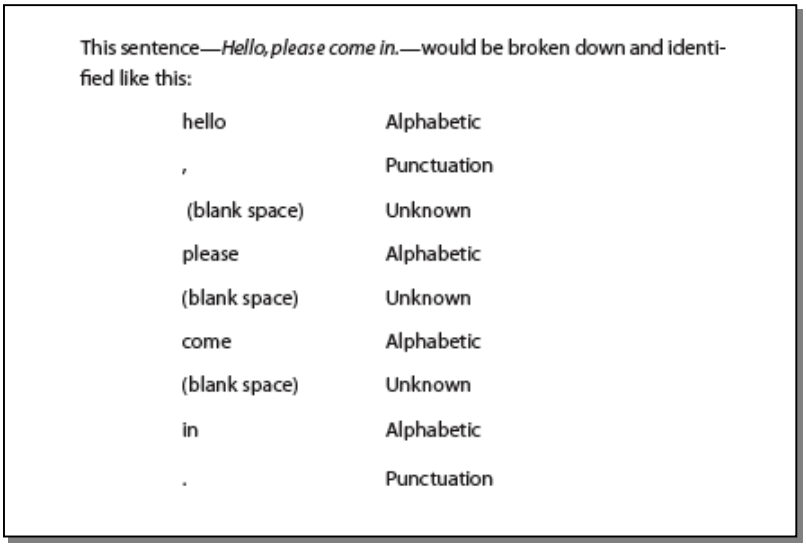
Les types de tokens de Verity sont (Verity, 2001b, pp. 7-8, [75]) :

⇒ alphabétique (« alphabetic »)
⇒ ponctuation (« punctuation »)
⇒ abréviations (« abbrevations »)
⇒ inconnu (« unknown »)

Les fins de paragraphe et de phrase dépendent des locales, qui utilisent en priorité la ponctuation. Pour les langues européennes, les marques les plus utilisées sont : le point (.), le point d'exclamation (!), le point d'interrogation (?) et les points de suspension (...).

Les abréviations sont traitées à l'aide d'un dictionnaire d'abréviations pour chaque langue (ex. en anglais Mr. pour Mister). Les chiffres, les dates et les pourcentages sont considérés comme inconnus.

Les traits d'union ne sont pas considérés à priori comme des séparateurs. Les mots contenant des traits d'union sont marqués inconnus et traités à une phase ultérieure. Si 1 token sur 10 est inconnu, les temps de traitement additionnels lors de l'indexation sont négligeables. Pour le traitement du français, les traits d'union entre les verbes et les pronoms (ex. : « donne-le-moi ») sont séparés, ainsi que les élisions (« l'abri »).



This sentence—*Hello, please come in.*—would be broken down and identified like this:

hello	Alphabetic
,	Punctuation
(blank space)	Unknown
please	Alphabetic
(blank space)	Unknown
come	Alphabetic
(blank space)	Unknown
in	Alphabetic
.	Punctuation

Source : : Verity Internationalization : Enabling E-business in Multiple Languages : Verity White Paper, Verity, 2001, p. 9, (Verity, 2001b, [75])

Figure 13 - Exemple de tokenization par Verity

❑ Paramétrage du découpage à l'installation de Verity K2

Pour la version 4.0.1 du moteur K2, il est possible de paramétrer (en lignes de commande, lors de l'installation de l'application) la recherche portant sur des caractères non-alphanumériques. Ceux-ci doivent être déclarés dans un fichier dont la syntaxe est expliquée dans la documentation de l'éditeur (Verity, 2001a, p. 7-4, [77] ; Verity, 2001c, p. 8-11, [73]). Cela permet à l'analyseur lexical (« lexical

analyser » ou « lexer ») de les prendre en compte lors de l'indexation des collections et lors de la recherche.

En standard, pour un traitement basique, le paramétrage de la segmentation des paragraphes et des phrases se fait en choisissant une des deux options : PSW (Paragraph-Sentence-Word, qui effectue un découpage basé sur la ponctuation) ou WCT (Word-Count, qui effectue un découpage statistique : 15 mots pour une phrase et 100 pour un paragraphe). La dernière est l'option par défaut. Quelle que soit l'option prise, les opérateurs de proximité <Phrase> et <Sentence> restent opérationnels, mais dans le deuxième cas, uniquement la distance entre deux mots, mesurée de façon dynamique, est prise en compte, sans tenir compte de la ponctuation (Verity, 2001a, p. 7-17, [77] ; Verity, 2001c, p. 8-7, [73]).

Si un module de traitement linguistique (une « locale ») a été installé, comme le français, le comportement du tokenizer est défini par la locale en fonction des particularités de la langue traitée et ne peut pas être modifié.

La casse des mots est prise en compte par défaut à l'indexation, donc, pour permettre une recherche insensible à la casse des mots, un paramètre est à changer en ligne de commande lors de l'installation. Ainsi, l'index construira des équivalences (Verity, 2001a, p. 7-17, [77] ; Verity, 2001c, p. 8-7, [73]).

1.3.3.1.2 LEMMATISATION / STEMMING

Pour permettre de retrouver des documents dans lesquels apparaissent différentes formes du même mot, les techniques qui peuvent être appliquées sont : le stemming (plus adapté à l'anglais) et la lemmatisation (plus adaptée au français).

⇒ STEMMING (DERIVATIONAL STEMMING)

« Derivational stemming », traduit en français par racinisation ou dérivation inverse, traite les formes de dérivation lexicale : tous les mots dérivés d'une même racine sont ramenés à leur racine, sans différence de classe grammaticale. Ainsi, on ramène à la même racine, « tomb* », les occurrences de « tombe », « tombeur », « tomberaient », « retomberas », « tombant » ; et à la racine « montr* » le verbe « montrer » et le nom « montre ». Certains systèmes procèdent ensuite à une désambiguïsation en fonction de la syntaxe de la phrase. Le problème de cette technique est le bruit : des mots de sens assez éloigné peuvent être regroupés sous la même racine et engendrer un bruit considérable lors de la recherche (Lefèvre, 2000, p. 38, [4]).

Pour le français, la situation est plus complexe que pour les langues à grammaire minimale et présentant un nombre réduit d'exceptions, comme l'anglais (Savoy, 1993, [78] ; Pierrel, 2000, p. 525, [53], chapitre rédigé par F. Namer) et, à ce titre, la lemmatisation est jugée préférable au stemming pour traiter des corpus de langue française.

⇒ LEMMATISATION (INFLECTIONAL STEMMING)

La lemmatisation est une opération qui consiste à faire correspondre la forme rencontrée dans le texte, à un « lemme », c'est-à-dire l'unité grammaticale minimale qui correspond, en général, à l'entrée d'un dictionnaire. Par exemple, toutes les formes d'un verbe seront ramenées à son infinitif (ex. : j'irai → aller), les noms au pluriel sont ramenés au singulier et les formes de féminin et/ou pluriel d'un adjectif sont ramenées au masculin singulier.

Ce traitement de la dérivation flexionnelle (des formes fléchies) peut s'appuyer sur des dictionnaires électroniques ou s'en passer.

Dans le cas de l'utilisation d'un dictionnaire électronique, le problème à résoudre est l'apparition d'un « mot inconnu », c'est-à-dire non listé dans le dictionnaire. Un module de prédiction (« guesser » en anglais) basé sur des règles peut alors résoudre une partie des ambiguïtés. Les règles sont du type de la généralisation morphologique : par exemple, on considère que toute forme se terminant par « ment » est un adverbe (Pierrel, 2000, p. 135, [53], chapitre rédigé par P. Paroubek et M. Rajman).

Dans le cas d'un système sans dictionnaire, tous les mots sont considérés à priori comme inconnus et le système s'appuie uniquement sur des règles (avec ou sans apprentissage) ou sur des méthodes probabilistes (traitements d'ordre statistique).

- **La lemmatisation et Verity K2**

Pour K2 de Verity, la lemmatisation est un traitement apporté par les modules additionnels de traitement des langues, à travers l'installation d'une ou plusieurs des 26 locales.

En l'absence de locale pour une langue particulière, le moteur peut la traiter à bas niveau, sans dictionnaire, en indexation full-text, pour peu que le jeu de caractères soit reconnu.

Selon l'éditeur, les 26 locales, qui permettent la lemmatisation, s'appuient sur des dictionnaires de référence des langues traitées. L'équipe commerciale de l'éditeur affirme que, en ce qui concerne le français, les traitements linguistiques sont basés sur le Petit Robert[25].

1.3.3.1.3 ÉTIQUETAGE (TAGGING)

L'étiquetage est effectué par un programme appelé étiqueteur (« tagger » en anglais) Habert, 1997, p. 165, [46]) :

> « *Attribuer à chaque mot la ou les étiquette(s) possible(s) peut se faire par consultation d'un dictionnaire, où chaque forme est suivie d'une liste de catégories,*

[25] Le Petit Robert contient plus de 60.000 entrées et, sur support électronique (CD-Rom édité en 2001), 140.000 formes fléchies, mais n'est pas, à proprement parler, un dictionnaire électronique, comme ceux élaborés par le LADL (Silberztein, 1993, [54]), ou ceux commercialisés par la société Memodata de Caen (Memodata, 1999, [55]). (voir à ce sujet partie 1.3.1.2.1 LA LINGUISTIQUE, page 41, et la note no. 13, page 43)

soit par analyse morphologique, soit par combinaison des deux techniques. Pour lever l'ambiguïté, deux solutions, qui peuvent d'ailleurs être associées, s'offrent alors : le recours à des règles ou l'appel aux probabilités [...]. »

À chaque mot sera ainsi attribuée une étiquette ou plusieurs. Dans ce dernier cas, la désambiguïsation consistera à choisir une seule étiquette parmi les possibilités listées.

- **Verity et l'étiquetage**

L'analyseur morphologique (« morphological analyser ») utilise des dictionnaires et interprète chaque token de type alphabétique en lui assignant une valeur, comme dans l'exemple suivant. (voir ci-dessous Figure 14 - Étiquetage effectué par Verity (Part-of-speech Tagging)).

As an example, the tagger would return these results for the following sentence: *When we speak, our words carry any number of messages just because of the tone our voices assume.*

Token	Tag
When	Subordinating conjunction
we	Pronoun
speak	Verb-present
,	Punctuation, comma
our	Determiner-possessive
words	Noun-plural
carry	Verb-present
any	Determiner-singular
number	Noun-singular
of	Preposition of
messages	Noun-plural
just	Adverb
because of	Preposition
the	Definite determiner
tone	Noun-singular
our	Determiner-possessive
voices	Noun-plural
assume	Verb-present
.	Sentence-ending punctuation

Source : Verity Internationalization : Enabling E-business in Multiple Languages : Verity White Paper, Verity, 2001, p. 9, (Verity, 2001b, [75])

Figure 14 - Étiquetage effectué par Verity (Part-of-speech Tagging)

1.3.3.1.4 EXTRACTION DES GROUPES NOMINAUX (NOUN PHRASE EXTRACTION)

L'extraction des groupes nominaux est une phase importante dans la désambiguïsation en fonction du contexte. L'étape se situe au niveau de l'analyse syntaxique. Certains systèmes construisent des graphes complets des phrases afin de lever un maximum d'ambiguïtés, ce qui nécessite un temps de traitement assez long, ralentissant l'indexation.

Un groupe nominal (ou syntagme nominal, en anglais « noun phrase »), noté GN, est une structure syntaxique hiérarchisée autour d'un substantif (Fuchs et Victorri, dans Fuchs, 1993, p. 110, [44]) :

> *« Le **syntagme** est un groupe d'unités, dominé par une « tête » dont la catégorie donne son nom au syntagme (syntagme nominal, syntagme verbal, syntagme adjectival, etc.) qui occupe une certaine position sur la chaîne, et joue un rôle fonctionnel donné. »*

Un syntagme constitué uniquement des catégories obligatoires est dit minimal.

> *« Le syntagme est dit **minimal** lorsqu'il n'est constitué que des catégories obligatoires qui le définissent (ainsi un déterminant + un nom pour le syntagme nominal, ex. :* le cordon*) ; il peut, au contraire comporter des ajouts facultatifs (ainsi, un adjectif, un groupe prépositionnel ou une relative, qualifiant le nom ; ex. :* le cordon gris *;* le cordon le plus ancien *;* le cordon qui relie l'ordinateur à l'imprimante*).*

Les groupes nominaux qui intéressent le plus sont les expressions figées (syntagmes lexicalisées). Leur traitement se situe donc aux frontières entre la morphologie et la syntaxe, car le cas des mots composés relève normalement de la morphologie. Mais, pour le traitement automatique, un mot composé n'a pas forcément de trait distinctif (ex. : tiret ou apostrophe ; quoique ambigus eux aussi, voir partie 1.3.3.1.1 SEGMENTATION (DÉCOUPAGE, TOKENIZATION), page 64) et peut être considéré par le tokenizer comme étant deux mots distincts.

Concernant les mots composés (qui, rappelons-le à nouveau, relèvent de la morphologie), pour le traitement du français, les types de mots composés ont été étudiés et il existe plusieurs types de structures, déterminées par l'équipe du LADL (Silberztein, 1993, [54] ; Fuchs, 1993, p. 93, [44] ; Lefèvre, 2000, p. 34, [4]), comme, par exemple :

- ⇒ nom + de + nom (« N de N ») : « pomme de terre », « mur du son »
- ⇒ nom + à + verbe (« N à V ») : « machine à laver », « pâte à crêpes »
- ⇒ nom + nom (« N N ») : « bateau mouche », « assurance-chômage »
- ⇒ nom + adjectif (« N A ») : « chaise longue », « carte bleue »
- ⇒ adjectif + nom (« A N ») : « long métrage », « grand-mère »
- ⇒ verbe + nom (« V N ») : « couvre-feu », « remue-ménage »
- ⇒ préposition + nom (« P N ») : « sans-abri », « sous-développement »

L'extraction des groupes nominaux a donc pour but de repérer ces structures porteuses de sens.

- **Verity et l'extraction des GN**

L'extraction des groupes nominaux est différente selon les langues. Elle est gérée par chaque « locale » installée, en fonction de règles spécifiques (Verity, 2001b, p. 10, [75]), afin d'identifier les notions uniques exprimées par des plusieurs mots.

> *« Noun Phrase extraction identifies concepts and physical entities that are described by more than one word. »*

Pour l'anglais, la définition d'un groupe nominal (Verity, 2001b, p. 10, [75]) est simple :

> *« The definition of a noun phrase varies somewhat from language to language, but the English definition is typical:*
> *-Noun phrases or proper noun phrases with their modifiers, including adjectives and adverbs : very large company, White House, biggest problem.*
> *-Noun phrases containing prepositional phrases that start with "of" : name of the game. »*

L'extraction des groupes nominaux, qui sont organisés selon des structures spécifiques à chaque langue, est importante pour permettre la recherche par catégories.

1.3.3.1.5 ÉLIMINATION DES MOTS VIDES (STOP LIST)

Pour l'indexation d'un texte en vue de la recherche d'information, tous les mots ne sont pas forcément significatifs (ex. : le, de). Ils peuvent être exclus de la recherche en les incluant dans un « dictionnaire des mots vides » appelé aussi « antidictionnaire » (« stop list » en anglais). L'avantage majeur est l'augmentation de la vitesse d'indexation et de recherche. Le désavantage est qu'elle peut contenir tout de même des mots utiles à la recherche.

- **Paramétrer une « stop list » pour Verity K2**

Si le besoin se fait sentir, il est possible de constituer un tel « dictionnaire des mots vides » en utilisant le fichier approprié (Verity, 2001a, p. 7-9, [77] ; Verity, 2001c, p. 8-3, [73]). Il s'agit d'un fichier ASCII à plat, et les mots à exclure sont à entrer en ligne de commande. Il est sensible à la casse des mots, donc, pour un même mot, il faut entrer trois orthographes : minuscules, majuscules, majuscule en début de phrase.

1.3.3.1.6 FILTRAGE PAR CONCEPTS (TOPIC SET)

Le filtrage par concepts exploite des relations lexicales sémantiques à travers des graphes de concepts (Pierrel, 2000, p. 246, [53], chapitre rédigé par C. Fluhr). Le graphe doit avoir été construit et le fonds indexé en l'utilisant, afin de s'en servir pour l'expansion de la requête de l'utilisateur.

Cette méthode de filtrage (par concepts ou topics) diminue le bruit engendré par d'autres types d'expansion de requêtes, comme, par exemple, la recherche floue (Lefèvre, 2000, p. 152, [4]).

- **Verity et le filtrage par concepts**

Le filtrage par concepts est transparent pour l'utilisateur final, si le choix a été fait d'interpréter la requête de l'utilisateur par les parsers Simple Query ou Free Text (Verity, 2001e, p. A-4 – A-6, [74]). Si les mots de la requête (dans le premier cas séparés par des virgules, dans le second, « langage naturel », 128 caractères maximum, séparés par des blancs) correspondent au nom de concepts définis dans le graphe de concepts (appelé « Topic Set »), le concept (« Topic ») correspondant sera utilisé pour rechercher les documents, mais aussi tous les concepts subordonnés[26]. Les résultats dépendent de la qualité du graphe de concepts construit dans l'application, par le service qui met en place le moteur.

L'offre standard de l'éditeur comprend un module de gestion des concepts (et de la taxonomie), appelé Verity Intelligent Classifier. Sa description sera abordée dans la partie 1.3.3.2 INTELLIGENT CLASSIFIER, MODULE DE GESTION DES CONCEPTS ET DE LA TAXONOMIE, page 75). La construction des concepts (Topics) se fait à l'aide de cet outil (voir partie 1.3.3.2.2 TOPIC SET (CONCEPTS), page 76).

Afin d'obtenir les meilleurs résultats lors de la construction du graphe de concepts (Topic Set), il convient d'utiliser une méthode de construction basée sur des critères logiques (voir partie 1.2.3 DÉMARCHES DE CONSTRUCTION, page 28).

1.3.3.1.7 CATÉGORISATION / CLASSIFICATION AUTOMATIQUE (AUTOMATIC CLASSIFICATION / CATEGORIZATION)

Les deux notions ne sont pas synonymes, mais ont ceci en commun : il s'agit de rapprocher les documents similaires en classes ou catégories, ce qui facilite une recherche thématique pour l'utilisateur final.

La catégorisation automatique et la classification automatique sont rapidement définies ainsi (Chaudiron, 2001, p. 198, [58]), évoquées comme fonctionnalités dans le contexte du text mining :

> *« organisation des documents du corpus par thèmes, en mode supervisé (catégorisation) ou non-supervisé (classification) »*

[26] C'est l'équivalent de l'autopostage descendant dans les thésaurus.

Les deux notions sont quelquefois utilisées l'une à la place de l'autre, selon les éditeurs, prévient Catherine Leloup (Leloup, 2002, [79]), qui fournit les définitions suivantes :

⇒ CATÉGORISATION AUTOMATIQUE

La catégorisation est définie (Leloup, 2002, [79]) comme :

> *« [...] aptitude à classer un document selon une ontologie prédéfinie – plan de classement, répertoires, taxinomie,... [...] »*

Ce qui est automatique, dans ce cas, est l'utilisation de ce plan de classement (ou taxonomie) prédéfini lors de l'indexation du fonds par le moteur de recherche et l'attribution, en fonction de l'algorithme utilisé par l'éditeur, de classes (catégories) prédéfinies à chaque document du corpus. Une ou plusieurs catégories peuvent être attribuées à chaque document.

⇒ CLASSIFICATION AUTOMATIQUE

La classification est définie (Leloup, 2002, [79]) ainsi :

> *« [...] aptitude à regrouper des documents en classes sans ontologie externe [...] »*

Le processus est automatique, exécuté selon des algorithmes. Il est possible de choisir un nombre de classes souhaité (mode supervisé) ou non (non supervisé).

La fonctionnalité de classification automatique des documents indexés par un moteur linguistique a pour base la catégorisation préalable manuelle ou automatique (en mode supervisé ou non-supervisé). La catégorisation entièrement automatique est appelée classification automatique.

Les moteurs statistiques effectuent uniquement une classification automatique à partir des documents, sans construction préalable de taxonomie.

La position prise ici sera la suivante : pour pouvoir bénéficier de la fonction appelée classification automatique des documents, il est nécessaire d'abord de procéder à la catégorisation des documents. La catégorisation (construction des catégories = taxonomie) peut être soit automatique (en mode supervisé ou non-supervisé), soit manuelle, par l'édition de catégories basées sur des règles et des concepts.

❑ Verity et la catégorisation / classification automatique

Concernant le moteur de recherche Verity, les possibilités de construction de taxonomies offertes par l'interface de gestion des concepts et de la taxonomie sont multiples. Toutes ces possibilités sont décrites dans le manuel d'utilisation du module Intelligent Classifier (Verity, 2001d, [80]). Elles sont traitées dans la partie 1.3.3.2 INTELLIGENT CLASSIFIER, MODULE DE GESTION DES CONCEPTS ET DE LA TAXONOMIE, page 75, plus précisément, dans la sous-partie 1.3.3.2.3 TAXONOMY (TAXONOMIE), page 78. Des livres blancs de l'éditeur, librement accessibles en ligne, permettent de se faire une petite idée des fonctionnalités disponibles (Verity, 2002, [81] ; Verity, 2003, [82]) et de l'apparence de l'interface de gestion.

1.3.3.1.8 RÉSUMÉ AUTOMATIQUE (AUTOMATIC SUMMARIZATION)

Le résumé automatique est une fonction qui permet de visualiser rapidement dans la liste de résultats quelques phrases du texte. Le filtrage sémantique vise à enrichir cette fonction par des traitements sémantiques en repérant les structures qui donnent des indications sur la teneur du discours.

Les méthodes de résumé automatique sont de plusieurs types (Minel, 2002, pp. 17-20 [57] ; Kayser, 2001, [56] ; Chaudiron, 2001, [58] ; Pierrel, 2000, pp. 255-266, [53] chapitre rédigé par J.-P. Desclès et J.-L. Minel) :

⇒ Méthodes fondées sur la compréhension

Sont basées sur la construction d'une représentation du texte, quelquefois fondée sur une analyse syntaxique. Cette méthode se fonde sur l'hypothèse implicite que le texte est bien écrit, or, cette hypothèse est trop forte. Les outils de ce type n'ont pas atteint le stade de l'industrialisation.

⇒ Méthodes par extraction

Partant de l'idée qu'il existe dans tout texte des unités saillantes, elle consiste à extraire ces unités en respectant l'ordre dans lequel elles apparaissent dans le texte. Les différents algorithmes sont basés soit sur le calcul de score en fonction de la fréquence des mots dans le texte, soit sur un calcul de similarité, paragraphe par paragraphe, soit sur le repérage de phrases prototypiques (« cue-phrases » en anglais) en se basant sur des mots ou expressions présentes dans le texte (comme, par exemple, « notre travail », « présenté précédemment »), qui offrent des indications concernant l'importance des phrases.

⇒ Méthodes par filtrage sémantique

Basée sur la « méthode d'exploration textuelle », cette méthode identifie les ressources linguistiques en les resituant dans leur contexte et en les organisant en fonction de tâches spécifiques. Les règles d'exploration contextuelle sont ensuite définies par le linguiste afin d'attribuer des étiquettes sémantiques.

Dans les applications commercialisées, le seul type employé pour l'instant est le résumé automatique basé sur l'extraction des phrases qui contiennent des termes dont le poids dans le document est calculé par des méthodes statistiques.

- **Le résumé automatique de Verity K2**

Le résumé automatique affiché avec la liste des résultats par les applications qui utilisent le moteur de recherche Verity K2 Enterprise est fondé sur l'extraction de phrases.

Le type de résumé et sa longueur sont paramétrables à l'installation (Verity, 2001c, pp. 8-6 – 8-7, [73]). Trois types de résumés sont possibles : les « meilleures » *n* phrases (celles qui obtiennent le meilleur score), les *n* premières phrases et les *n* premiers caractères ; où *n* représente une variable paramétrable en fonction des besoins de l'application (longueur du résumé).

1.3.3.1.9 INVERSION

L'opération consiste à constituer un (des) index positionnel(s) inversé(s), qui stocke(nt) la position du mot dans le document et la clé du document (voir Figure 12 - Opérations effectuées par le moteur Verity (exemple d'application multilingue), page 63). Le moteur d'indexation de Verity indexe automatiquement les documents, selon les configurations définies lors des différents paramétrages.

On obtient ainsi les collections, constituées d'index des mots et de tables des documents. (Verity, 2001a, p. 1-7, [77]) :

> *« A collection consist of word indexes, document tables, and optional indexes used for specialized functions. Each collection is subdivised into units called partitions, and for each partition there is a word index and documents table. »*

Les documents sont laissés à leur place et le moteur conserve uniquement ses fichiers en format propriétaire, sous la forme de tables VDB (Verity Databases)

C'est une opération d'ordre informatique, qui a pour but de stocker toutes les informations obtenues lors des analyses précédentes afin de les exploiter lors de la recherche.

1.3.3.2 INTELLIGENT CLASSIFIER, MODULE DE GESTION DES CONCEPTS ET DE LA TAXONOMIE

Verity Intelligent Classifier (VIC), interface Windows dédiée à la gestion des concepts et de la taxonomie, est une alternative conviviale à la gestion au moyen de lignes de commande, qui reste cependant possible (voir Figure 15 - Interface de l'outil Verity Intelligent Classifier, page 75). La représentation graphique est plutôt claire et parlante, d'un maniement assez intuitif. La possibilité de glisser-déposer (Drag and Drop) est agréable pour faire des changements, mais peut s'avérer dangereuse lorsqu'on effectue des mises à jour, car la possibilité de créer une boucle[27] en déplaçant les fragments n'est pas à négliger.

1.3.3.2.1 INTERFACE

L'interface de Verity Intelligent Classifier se présente ainsi :

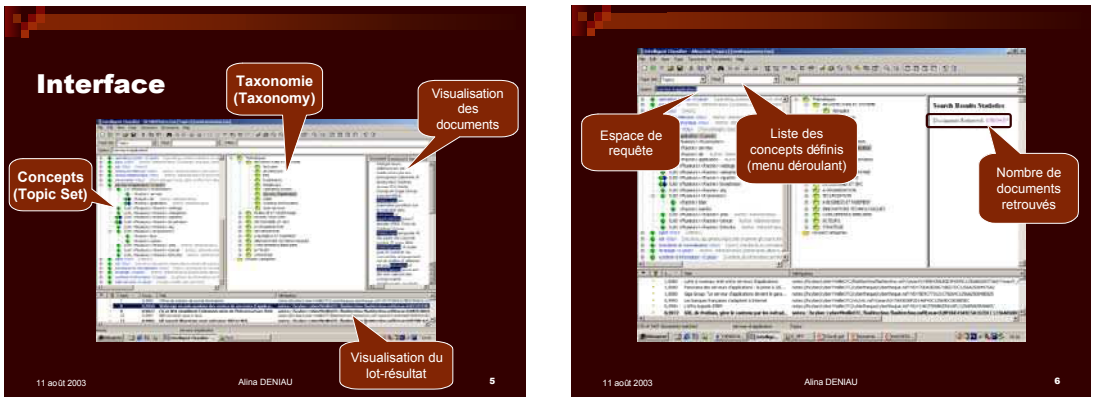


Figure 15 - Interface de l'outil Verity Intelligent Classifier

Les quatre panneaux peuvent être affichés tous en même temps ou fermés un par un pour en privilégier un aspect de la gestion (Topics ou Taxonomy).

Il est possible de faire des requêtes en mode expert afin d'afficher les résultats d'interrogation afin de tester leur pertinence. Les documents du lot-résultat peuvent être visualisés un par un et les mots de la requête sont mis en surbrillance dans l'espace de visualisation.

L'interface et ses fonctionnalités sont présentées dans le guide d'utilisation fourni par l'éditeur (Verity, 2001d, [80]).

[27] On obtient une boucle lorsqu'un concept père est déplacé ou copié de telle façon qu'il se retrouve dans la hiérarchie aussi comme fils de son fils (*père — fils — père). Rappelons que la seule construction valable est : **grand-père — père — fils**. Lorsqu'on crée les concepts, le parseur empêche de créer des boucles, mais lors des déplacements de concepts déjà créés, cela peut arriver par mégarde (par exemple en ayant lâché trop tôt le bouton gauche de la souris) ou par méconnaissance des principes de base.

1.3.3.2.2 TOPIC SET (CONCEPTS)

Les « concepts » ou Topics (dans la terminologie de Verity) sont des mots ou des groupes de mots (GN ou GV) qui expriment des sujets d'ordre général. Ils sont définis par des règles, sorte de requêtes préenregistrées, à l'aide des opérateurs Verity (langage d'interrogation propriétaire appelé Verity Query Language ou VQL) (voir partie 1.3.3.3 OPÉRATEURS, page 80, et l'annexe OPÉRATEURS VERITY, page 150) (Verity, 2001f, p. 1-4, [83]) :

> *« A topic is a grouping of information that comprises a topic definition related to a concept or a subject area. In other words, a topic is a save query which is used to search over a collection. »*

Un Topic Set est une organisation sous forme de graphe de concepts utilisable dans une application Verity (Figure 16 - Propriétés du Topic Set (Concepts)77. Le fichier obtenu à l'aide du module Intelligent Classifier doit être exporté dans l'application Verity pour pouvoir être pris en compte :

> *« A topic set is a grouping of topic definitions that have been compiled for use by a Verity application. This is where one or more topics is saved for searching on a collection. »*

Un concept est formé de trois niveaux :

- **Top-level topics (=têtes de hiérarchie)**

Ce sont des sujets d'ordre général. Ils se positionnent au premier niveau et apparaissent en bleu à l'affichage dans l'interface IC. Ce sont des « concepts », et ont obligatoirement un niveau inférieur.

- **Subtopics (facultatifs)**

Plusieurs sub-niveaux sont possibles. Ils sont des « concepts » aussi et se présentent aussi en bleu. Ils doivent être « définis » par un niveau inférieur. Ils peuvent être utilisés dans la gestion des synonymes.

- **Evidence topics (instanciations)**

Combinaison de caractères alphanumériques (max. 128). C'est le niveau le plus bas. Ils s'affichent en gris et ne sont pas des concepts, mais des instanciations qui servent à les définir.

Les noms des concepts peuvent être choisis librement, mais, faire correspondre le nom choisi avec les requêtes les plus vraisemblables facilite la recherche pour l'utilisateur final.

> *« The name of the topic is arbitrary, unless you want to replace topics with user queries in which case you must be able to determine names that users are likely to type as queries. »*

Dans ce cas, le nom du concept formé de plusieurs mots (pluriterme) doit contenir des traits d'union à la place des espaces :

> « *When spaces are replaced by hyphens in a query, the application also uses a topic if the result matches a topic name. For example, if* general-motors *is the name of a topic and the user enters the query* general motors*, then the application returns the results of the topic.* »

Pour simplifier, en interrogation, l'utilisateur peut taper des mots ou groupes de mots figurant dans les concepts (mots en bleu) pour obtenir tous les textes dans lesquels figurent les chaînes de caractères (mots en gris) qu'on leur a attachées à l'aide des opérateurs.

> « *Verity K2 Enterprise interprets users' queries through the Verity Query Language (VQL). If a topic set is used, and a user types a search term that matches the name of a topic, then the application uses that topic instead of the search term. For example, if* products *is the name of a topic and the user enters the query products, then the application returns the results of the topic instead of the results that would be obtained by searching for the word* "products". »

Figure 16 - Propriétés du Topic Set (Concepts)

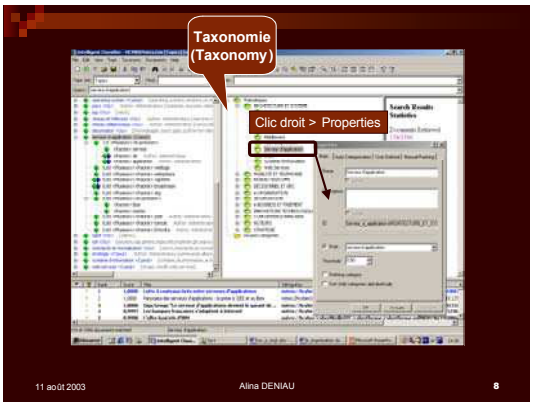


Figure 17 - Propriétés de Taxonomy (Taxonomie)

1.3.3.2.3 TAXONOMY (TAXONOMIE)

La construction de la taxonomie a pour but de permettre une recherche par catégories. Les catégories apparaissent dans l'application comme un système de répertoires et l'utilisateur final, en cliquant sur les dossiers, a accès à des catégories de plus en plus spécialisées.

Dans l'outil Inteligent Classifier, la « Taxonomy » (taxonomie) se présente sous la forme d'un arbre où toutes les catégories visibles dans l'application vers laquelle ils seront envoyés dépendent d'une seule racine (« Root ») (voir Figure 17 - Propriétés de Taxonomy (Taxonomie), page 77).

Les catégories sont caractérisées par un nom et un identifiant (unique), assimilé par l'éditeur aux langages contrôlés (Verity, 2001d, p. 5-3, [80]) :

> « *Each category must have a unique ID that identifies it within the taxonomy. The category ID is used for specifying categories for browsing and scoping in the browse server. It is also used for specifying categories assigned to documents when populating the Knowledge Tree, such as in document meta-tags, collection fields, or manual submissions to the Knowledge Tree.* ***In the library science sense, the set of category IDs represents a controlled vocabulary for the categorization and indexing of documents.*** »

Les catégories peuvent être construites de façon automatique (non-supervisée ou supervisée) ou manuelle (voir partie 1.3.3.1.7 CATÉGORISATION / CLASSIFICATION AUTOMATIQUE (AUTOMATIC CLASSIFICATION / CATEGORIZATION), page 71).

Lors de la construction de la taxonomie en mode manuel, les catégories peuvent être basées sur un concept défini dans le « Topic Set » ou sur des règles d'inférence (Business Rule ou Classify Rule) définies pour chacune. Les catégories subordonnées à une catégorie définie dans l'arborescence ne sont pas forcément soumises au principe d'héritage de propriétés. Pour l'activer, il est nécessaire d'utiliser l'option « Refining category », ce qui a pour conséquence d'additionner la « business rule » de la catégorie subordonnée à la « business rule » de la catégorie supérieure.

En mode automatique (catégorisation automatique en mode supervisé ou non-supervisé), pour créer des catégories, plusieurs façons de faire sont possibles, parmi lesquelles : la création à partir des métadonnées et la création à partir de clusters de documents similaires (Verity, 2001d, pp. 6-25 – 6-30, [80]).

Les catégories peuvent être créées aussi à l'aide de la technologie propriétaire LRC (Logistic Regression Classifier), en mode automatique (classification automatique, « LRC Classification »), à partir de groupes de documents marqués pertinents/non-pertinents (Verity, 2001d, p. 6-23, [80]).

1.3.3.2.4 DÉMARCHES POSSIBLES (« TOPIC DESIGN STRATEGIES »)

Parmi les trois types de démarches de construction de vocabulaire contrôlé (voir partie 1.2.3 DÉMARCHES DE CONSTRUCTION, page 28), deux sont décrits succinctement dans la documentation fournie par l'éditeur (Verity, 2001d, pp. 2-3 – 2-4, [80]) pour la gestion des concepts (« Topic Design Strategy ») et de la taxonomie à l'aide Verity Intelligent Classifier.

La brièveté de la présentation présuppose la connaissance préalable :

- ⇒ des techniques de construction d'outils de contrôle de langage
- ⇒ des types de démarches possibles
- ⇒ des classes d'opérateurs employés par le langage de requête de Verity (VQL)

La maîtrise des techniques et des démarches (à choisir en fonction du contexte) s'avère nécessaire autant pour exploiter les possibilités de l'outil que pour aider les spécialistes du domaine à formaliser et structurer leur savoir et à hiérarchiser leurs connaissances, exprimées par les mots du métier, selon des critères méthodiques, explicites, formalisés et logiques.

Les deux démarches listées par l'éditeur sont :

- **« Top-Down Design »**

Cette démarche est conseillée par l'éditeur lorsqu'il s'agit de la gestion de fonds qui grandissent très rapidement (ce qui fait penser aux bases presse). Cette démarche de type classificatoire est utilisée notamment lors de la définition de plans de classement (voir partie 1.2.3.2.2 « TOP-DOWN » OU DÉMARCHE CLASSIFICATOIRE, page 36). La méthode est très brièvement décrite ainsi dans le manuel d'utilisation : commencer par établir des grandes catégories (niveau Top-level topics) et subdiviser ensuite en classes plus petites (niveau Subtopics) jusqu'aux termes « importants » du domaine et à son jargon (niveau Evidence topics).

- **« Bottom-Up Design »**

Conseillée par l'éditeur pour les corpus contenant beaucoup de documents similaires (ce qui fait penser aux corpus métier). Cette démarche est pratiquée dans la construction des thésaurus, avec de très bons résultats, car elle oblige à formaliser chaque critère de tri lors de la construction. Même si les liens ne sont pas étiquetés et ne sont pas forcément visibles dans l'application, lors de la phase de construction et de mise à jour, le tri par thèmes et par facettes permet d'obtenir une base de connaissances cohérente (voir partie 1.2.3.2.1 « BOTTOM-UP » OU DÉMARCHE THÉSAURUS, page 32). Le manuel l'expose très succinctement ainsi : à partir d'une liste de mots significatifs extraits des documents (Evidence topics), grouper successivement les mots en classes supérieures (Subtopics et Top-level topics).

1.3.3.3 OPÉRATEURS

Les opérateurs de Verity sont présentés dans deux guides fournis avec le moteur :

- Verity K2 Enterprise, Intelligent Classification Guide (Verity, 2001d, [80])
- Verity Query Language Guide (Verity, 2001f, [83])

D'une apparence complexe, le langage d'interrogation de Verity (Verity Query Language ou VQL) est un outil puissant qui permet de construire des requêtes complexes en mode expert. Utilisable autant en interrogation que pour la construction de concepts (Topics) basés sur des requêtes ou de la taxinomie (Taxonomy), le VQL utilise plus de 35 opérateurs.

Des règles de précédence sont à respecter dans l'usage des opérateurs. Un tableau synthétique est présenté dans le guide d'utilisation de Verity Intelligent Classifier (Verity, 2001d, p. 2-5, [80]). La syntaxe est validée par le parser lors de la création des concepts. Le parser refuse la création de concepts invalides et affiche des messages d'alerte et d'explication en cas de fausse manoeuvre.

1.3.3.3.1 TYPOLOGIE DES OPÉRATEURS EXISTANTS ET LEUR EXPRESSION CHEZ VERITY

Les types d'opérateurs utilisés par les langages de requêtes sont de plusieurs types, bien connus et utilisés depuis longtemps pour interroger les logiciels documentaires et les bases de données accessibles en ligne via des moteurs de recherche inclus (Lefèvre, 2000, pp. 147-152, [4]). Certains de ces opérateurs sont standards, d'autres plus spécifiques à certains logiciels.

- **Opérateurs booléens**

⇒ ET : ET logique (pour Verity : <All>)

⇒ OU : OU non exclusif (pour Verity : <Any>, en français <Quelconque>)

⇒ SAUF : exclusion de plusieurs mots ou concepts (pour Verity : <Not>, en français <Sauf>)

⇒ PARMI : agit comme un ET ou comme un OU (peu utilisé, intérêt limité)

- **Opérateurs booléens pondérés**

⇒ ET_PONDÉRÉ : place en tête les documents contenant le maximum de termes de la requête ; selon les variantes, il élimine ou non les documents qui ne contiennent pas tous les mots de la requête (pour Verity : <And>, en français <Et>)

⇒ OU_PONDÉRÉ : place en tête les documents qui possèdent de nombreuses fois les termes de la requête, et des termes à poids fort dans le corpus (pour Verity : <Or>, en français <Ou>)

⇒ OU_ET_PONDÉRÉ : donne en retour tous les documents qui contiennent au moins un des termes de la recherche spécifiés, mais positionne en tête ceux qui contiennent tous les termes à la fois, ou du moins ceux qui contiennent un maximum de termes (pour Verity : <Accrue>, en français <Cumul>)

❑ Opérateurs de proximité (et appartenance)

⇒ ADJACENCE : impose que les mots recherchés soient adjacents dans l'ordre donné (pour Verity : <Near>, en français <Proche>)

⇒ DISTANCE : recherche dans le texte un ensemble de mots éloignés au plus d'une distance n, sans notion d'ordre (pour Verity : <Near/n>, en français <Proche/n>)

⇒ ANTÉCÉDENCE : recherche dans le texte un ensemble de mots éloignés au plus d'une distance n, mais l'ordre des mots est pris en compte

⇒ PORTÉE : pour trouver un mot dans un intervalle précis, avant et après un mot donné

⇒ ENTRE : recherche un mot dans un intervalle de texte défini par deux autres mots

⇒ PRÉSENCE : sélectionne une association de mots présents dans une même portion de texte (pour Verity : <Phrase>, en français <Expression>)

⇒ PHRASE : impose que les mots recherchés soient à l'intérieur d'une même phrase (pour Verity : <Sentence>, en français <Phrase>)

⇒ PARAGRAPHE : impose que les mots recherchés se trouvent à l'intérieur d'un même paragraphe (pour Verity : <Paragraph>, en français, <Paragraphe>)

❑ Opérateurs prenant en charge les variantes des mots

⇒ TYPOGRAPHIE : joue sur la typographie, en acceptant ou non les substitutions majuscules-minuscules, ou les variantes d'accentuation (pour Verity : <Case>, en français <Casse>)

⇒ TRONCATURE (DROITE, GAUCHE, JOKER) : remplace une chaîne de caractères quelconque. La troncature peut être appliquée à gauche (en début de mot), au milieu, ou à droite (en fin de mot) (pour Verity : <Wildcard>, en français <Troncature> ; pour les symboles utilisés, voir tableau des opérateurs Verity en annexe, page 150)

⇒ MASQUE : réalise la même fonction que TRONCATURE, mais en remplacement ou en exclusion d'un seul caractère (pour Verity : <Wildcard>, en français <Troncature> ; pour les symboles utilisés, voir tableau des opérateurs Verity en annexe, page 150)

⇒ DÉRIVATION, DÉRIVATION INVERSE : étend la recherche à partir du mot donné, pris comme racine, à tous les mots obtenus par dérivation ; la dérivation inverse donnera la racine à partir d'une forme dérivée quelconque (pour Verity : <Stem>, en français <Racine>)

⇒ RESSEMBLANCE : fonction de recherche floue ou approximative sur les mots, qui admet des écarts sur un ou plusieurs caractères (pour Verity : <TypoEqual>, en français <Typo>)

⇒ PHONÉTISATION : recherche par ressemblance phonétique, lorsque l'orthographe est mal connue (pour Verity : <Soundex>, en français <Consonance>)

❑ Opérateurs sur les concepts et le domaine

⇒ EXTENSION DE SENS (SYNONYMIE, HIÉRARCHIE, ASSOCIATION) : étend la recherche à des mots des sens voisins (pour Verity, synonymie : <Thesaurus>, en français <Synonyme>)

⇒ RESTRICTION À UN SUJET : peut permettre d'organiser les informations, sur lesquelles on a préalablement appliqué un classement, dans un plan de classement (pour Verity peut être une case à cocher, comme dans l'exemple du portail Cyberthèque)

❑ Opérateurs de comparaison numérique

Ce sont des opérateurs applicables à des champs numériques ou date. Permettent, par exemple, de sélectionner des documents en fonction de leur date de création.

⇒ = (ÉGALE)
⇒ > (PLUS GRAND QUE)
⇒ < (PLUS PETIT QUE)
⇒ >= (PLUS GRAND OU ÉGAL À)
⇒ <= (PLUS PETIT OU ÉGAL À)

1.3.3.3.2 TYPOLOGIE DES OPÉRATEURS SELON LA TERMINOLOGIE VERITY

Un tableau qui reprend la terminologie Verity, explique leur usage et tente, dans certains cas, la traduction française, est placé en annexe page 150.

- **« Evidence operators »**

Sont à utiliser au niveau des instanciations (evidence topic).

- **« Field and Zone operators »**

Sont à utiliser sur les champs des bases de données et les zones des documents structurés.

- **« Proximity operators » (Opérateurs de proximité)**

- **« Combinatorial operators »**

Comprend la classe des opérateurs booléens et des opérateurs booléens pondérés.

- **« Behavior modifiers »**

Permettent d'agir sur le comportement des autres opérateurs selon des règles strictes de combinaison (ex. Casse, Ordre).

- **« Score operators »**

Permettent de rajouter des pondérations de type spécifique (ex. Sum, YesNo).

La présentation des systèmes d'organisation des connaissances, parmi lesquels les taxonomies, et la description du fonctionnement des moteurs de recherche, notamment K2 de Verity, seront illustrées par le cas concret de l'audit effectué lors de la mise à jour d'un langage contrôlé géré par le moteur de recherche Verity K2 dans un portail de veille technologique et concurrentielle d'une grande entreprise : la Cyberthèque de la Direction des systèmes d'information de la Société Générale.

DEUXIÈME PARTIE :

2 AUDIT : VERITY K2 (CONCEPTS ET TAXONOMIE) ET LE PORTAIL CYBERTHÈQUE

2.1 CONTEXTE ET DÉMARCHE PROJET

2.1.1 CONTEXTE

L'audit a été réalisé du 30 juin au 12 septembre 2003, dans le cadre d'un stage d'études encadré par la responsable du portail de veille technologique du pôle OTC, service de veille technologique de la Direction des systèmes d'information de la Banque de Détail de la Société Générale (INFO/STA/OTC).

❑ Le pôle OTC

Le pôle Opportunités Technologiques et Communication sur les Technologies (OTC) emploie au total neuf personnes dont une assistante de direction. La chef de service, les trois veilleurs et les deux ingénieurs études et expérimentations sont informaticiens de formation. Deux documentalistes ont en charge le centre de documentation informatique (fonds papier et catalogue informatisé accessible en ligne), situé dans un autre bâtiment.

OTC a pour missions de détecter, étudier et expérimenter les opportunités des nouvelles technologies pour évaluer l'intérêt business et la pertinence de l'usage de ces technologies dans le contexte de la banque de détail et anticiper leur intégration dans l'architecture globale du système d'information pour répondre aux exigences de la banque. Il contribue à la réflexion sur le potentiel des technologies dans la création de valeur pour la banque, en analysant les tendances du marché et les orientations techniques de la concurrence. Les activités d'OTC sont :

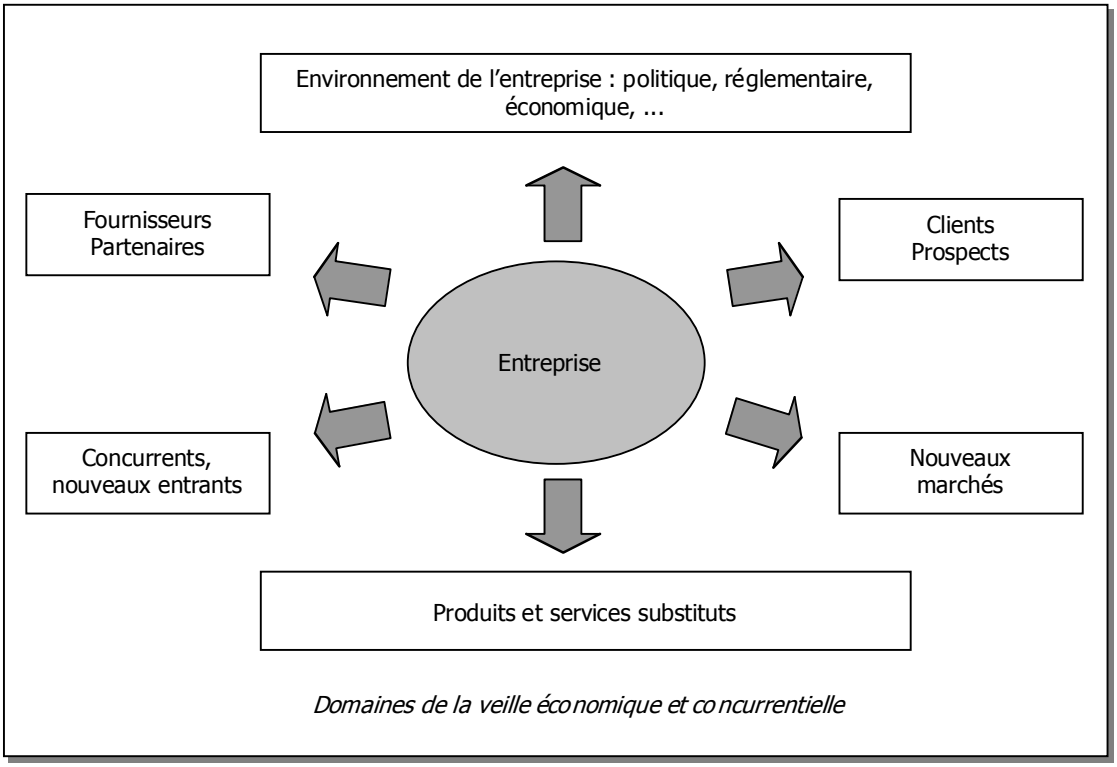
⇒ L'Observatoire : assurer le suivi et l'analyse des technologies de l'information, de la stratégie des fournisseurs et de l'usage de ces technologies par les acteurs du secteur banque-assurance pour recueillir et analyser les tendances de l'évolution technologique,

⇒ Les Études : réaliser des études sur l'état de l'art des technologies et leurs enjeux pour le secteur bancaire et analyser les opportunités que représentent ces technologies pour le business et le système d'information,

⇒ Les Expérimentations : conduire la réalisation de projets d'expérimentation de nouvelles technologies avec les directions bancaires intéressées, de manière à mettre en évidence la valeur ajoutée de la technologie pour le business, la pertinence de leur usage et contribuer concrètement aux décisions d'adoption et d'intégration de ces technologies par la banque,

⇒ La communication sur les technologies : contribuer à la compréhension par la Banque des Nouvelles Technologies de l'Information ainsi qu'à l'évaluation des opportunités qu'elles génèrent en diffusant l'information et en animant les comités de réflexion avec les acteurs des départements bancaires et informatiques.

L'observatoire des nouvelles technologies, cellule de veille sur les nouvelles technologies appliquées au domaine bancaire, effectue une veille technologique et concurrentielle, à partir de sources multiples. Les principaux domaines de la veille concurrentielle sont rappelés dans le schéma suivant :



Source : Catherine LELOUP, Moteurs d'indexation et de recherche, Eyrolles, 1998, p. 141 (Leloup, 1998, [43])

Figure 18 - Domaines de la veille concurrentielle

Dans le cadre des activités de veille, le portail Cyberthèque peut être considéré comme un outil de mutualisation des connaissances (Laviale, 1999, [84]). Il est accessible, depuis l'intranet groupe, à tous les employés disposant d'un poste informatique.

La Cyberthèque, portail de veille sur les nouvelles technologies appliquées au domaine bancaire, en est à sa troisième version depuis sa création (voir Figure 19 - La page d'accueil du portail Cyberthèque, page 88). La réalisation informatique des différentes versions est sous-traitée à une SSII.

Pour la partie technique, les documents accessibles via le portail Cyberthèque sont stockés dans plusieurs bases Lotus Notes (v.5.0.10). La charte graphique avait radicalement changé lors de la mise en place de la version deux du portail (mise en production en septembre 2001). La nouveauté de la troisième version est l'implantation du moteur de recherche K2 Enterprise de Verity (v.4.0.1) et elle a été mise en production en mars-avril 2002.

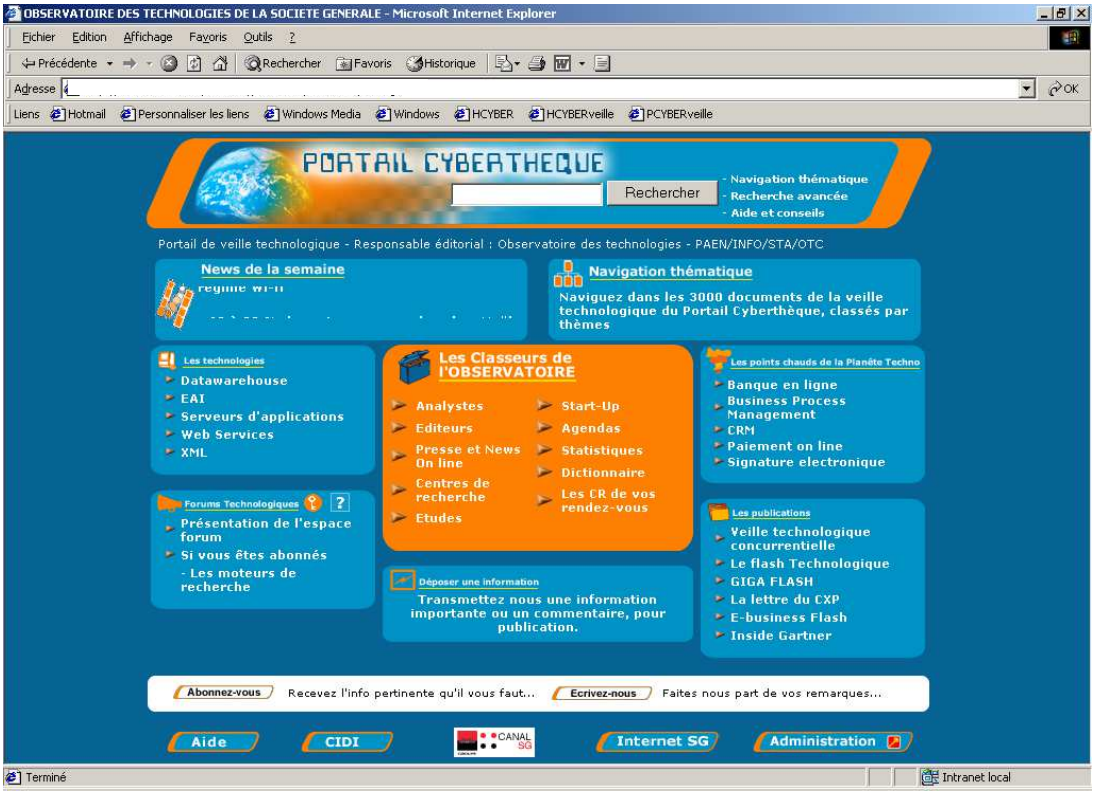


Figure 19 - La page d'accueil du portail Cyberthèque

Les documents accessibles via le portail de veille technologique et concurrentielle sont stockés dans plusieurs bases Notes. Les droits d'accès aux bases de la Cyberthèque sont gérés par la responsable du portail. Il existe quatre types de droits d'accès aux bases de la Cyberthèque :

⇒ lecteur (lecture, recherche, accès articles)
⇒ rédacteur (accès base, création articles, modification et suppression de ses propres articles)
⇒ éditeur (les mêmes + modification, suppression des articles des autres, validation des articles en vue de publication dans la Cyberthèque)
⇒ administrateur (tous les droits)

Les neuf personnes de OTC ont au minimum les droits de rédacteur et participent à des degrés divers à l'alimentation des différentes bases. La même personne peut avoir des droits différents selon les bases.

Actuellement la base Cyberthèque compte plus de 4800 documents, en anglais et en français, catégorisés et résumés automatiquement par le moteur de recherche K2 Enterprise de Verity.

❑ Acquisition du moteur de recherche K2 Enterprise de Verity

Afin de faciliter la recherche des documents internes accessibles au personnel de la Société Générale sur le portail Cyberthèque, l'équipe de l'Observatoire des nouvelles technologies a souhaité se doter d'un moteur de recherche.

Acquis suite à une étude de marché effectuée en juillet-août 2001 (Vacarie, 2001, [85]), le moteur de recherche K2 Enterprise (version 4.0.1) de la société Verity a été implanté dans le portail Cyberthèque en décembre 2001.

L'offre contient aussi l'outil de veille appelé spider Verity (crawler), qui indexe les pages des sites web déterminées à l'avance. Il utilise le même dictionnaire métier que le moteur de recherche.

L'offre standard de l'éditeur comprend l'outil de gestion de dictionnaires métier, Verity Intelligent Classifier, doté d'une interface graphique Microsoft Windows®, qui constitue une alternative plus conviviale à la gestion en mode console.

Le dictionnaire métier (concepts et taxonomie) a été mis en place entre janvier et mars 2002.

❑ Types de recherche

Outre la recherche simple, deux autres possibilités offertes par le moteur ont été mises en oeuvre : la recherche avancée, dont le formulaire de requête est très influencé par ceux disponibles dans les moteurs de recherche sur le Web (voir ci-dessous Figure 20 - La recherche avancée), et la recherche thématique, par navigation dans un plan de classement préétabli (la taxonomie).

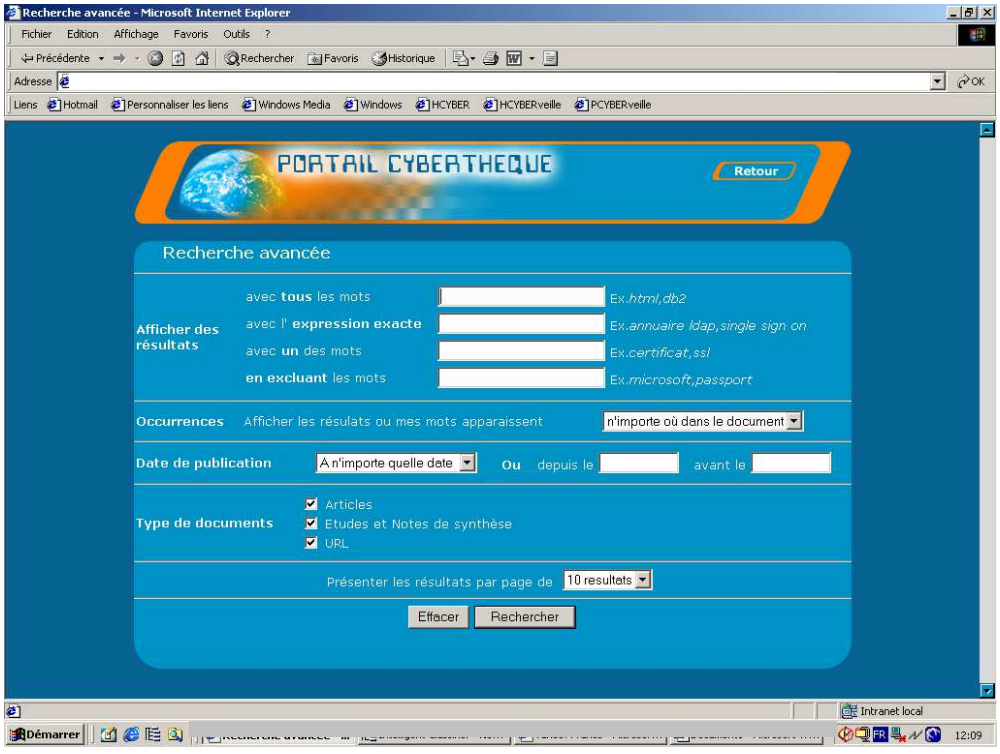


Figure 20 - La recherche avancée

❑ Version

La version du moteur de recherche Verity K2 Enterprise actuellement installée dans le portail Cyberthèque est la version 4.0.1. Une partie des fonctionnalités disponibles n'ont pas été mises en œuvre, n'ayant pas été jugées intéressantes à l'époque de l'étude de marché menée en juillet-août 2001.

❑ Évolution

La version du moteur actuellement commercialisée par la société Verity est la version 5.0. Pour l'instant, l'équipe OTC n'envisage pas l'évolution à la version supérieure. Toutefois, compte tenu des termes du contrat de maintenance et du rythme des mises à jour de la société éditrice, la migration vers une version supérieure sera certainement à examiner à plus ou moins court terme. Les différences fonctionnelles (une seule « collection » en indexation contre plusieurs actuellement) et les améliorations apportées au moteur, notamment concernant la gestion du multilinguisme (cross-language), rendent nécessaire la séparation de l'aspect intellectuel des concepts et leur formalisation dans l'outil. Cette contrainte a été prise en compte autant que possible tout au long de la mission de stage.

Une première approche de la problématique a été évoquée dès février 2003, lors de la première partie du stage. L'équipe OTC a fait état de ses attentes et de ses besoins concernant l'optimisation de la gestion du dictionnaire métier à l'aide de l'outil Intelligent Classifier de Verity (voir partie 2.2.1 ANALYSE DES BESOINS, page 95).

2.1.2 DÉMARCHE PROJET

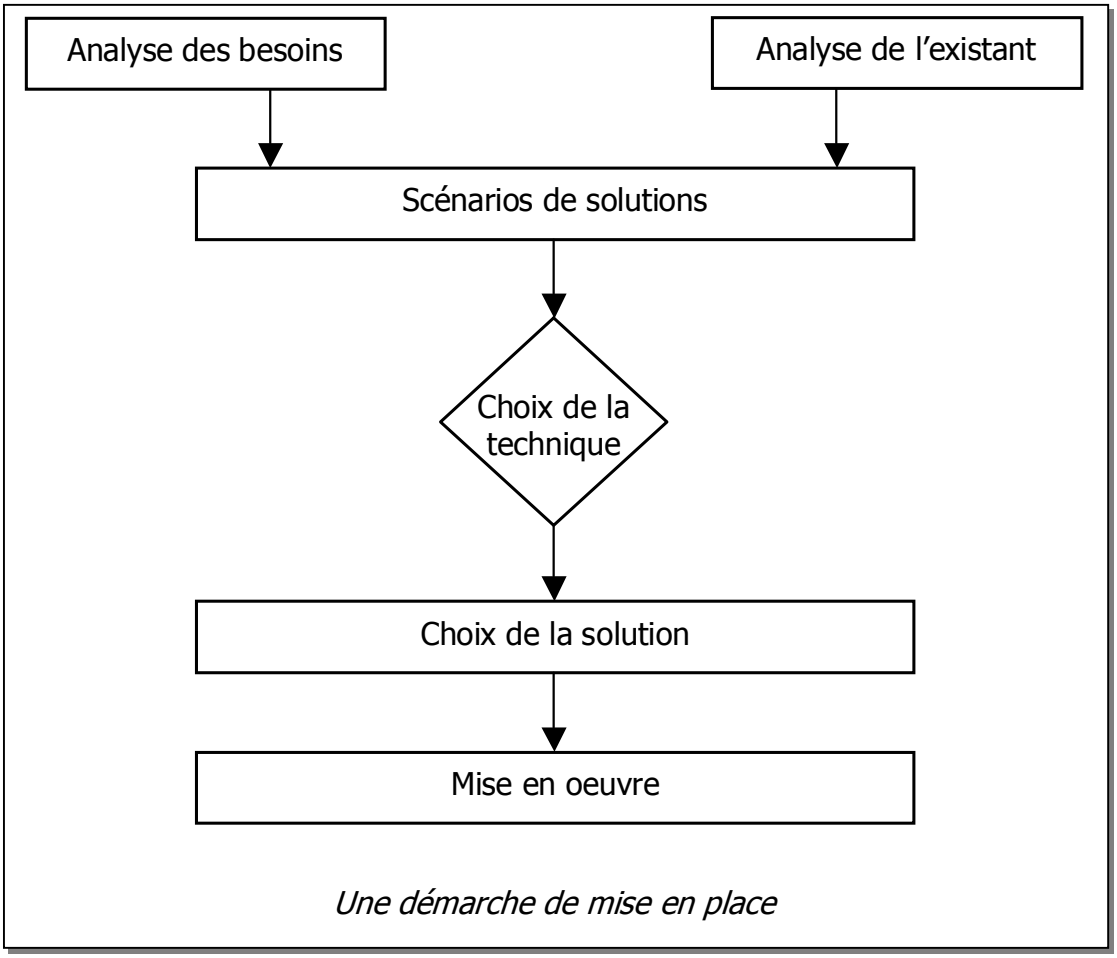
Malgré qu'il s'agisse d'une mission qui visait à améliorer la qualité d'un produit existant, les observations effectuées au mois de février 2003 ont déterminé le choix d'une démarche de type conduite de projet (Corbel, 2003, [86] ; Picq, 1999, [87]). La nécessité d'un audit de l'existant a aussi paru nécessaire (Morgat, 1995, [88]).

La mise en place du dictionnaire métier a été faite en l'absence d'une personne formée spécifiquement à la gestion de langage contrôlé, ce qui explique le manque de méthodologie et l'absence d'historique et de règles de gestion (documentation pour la mise à jour). Cette absence a eu des incidences lors de rajouts ponctuels de nouveaux termes, ce qui a nécessité une mise à plat de l'ensemble.

Un audit de l'existant s'est donc avéré indispensable pour deux raisons :

- ⇒ faire l'analyse de l'existant pour amener le changement
- ⇒ faire prendre conscience à l'équipe OTC du type de travail dont il est question et de la méthode/démarche à appliquer

Un schéma de démarche possible, emprunté à la consultante Catherine Leloup (Leloup, 1998, p. 149, [43]), montre les phases d'un projet (voir Figure 21 - Démarche projet (d'après Leloup, 1998), page 92). Bien que ce schéma illustre à l'origine la démarche conseillée pour aboutir au choix d'un moteur de recherche parmi les produits présents sur le marché, il se prête aussi au cas présent, dans la mesure où le contexte spécifique de la mission induit une réflexion de même ordre sur la gestion du langage contrôlé implanté dans le moteur Verity K2. Selon Catherine Leloup, en matière de moteurs de recherche, la meilleure solution est d'adapter fortement sa démarche aux cas rencontrés.



Source : Catherine LELOUP, Moteurs d'indexation et de recherche, Eyrolles, 1998, p. 149 (Leloup, 1998, [43])

Figure 21 - Démarche projet (d'après Leloup, 1998)

2.1.2.1 PHASE AUDIT

La phase d'évaluation a duré 5 semaines, du 7 juillet au 11 août 2003. Les détails de l'audit seront présentés dans la partie 2.2 AUDIT DES CONCEPTS ET DE LA TAXONOMIE, page 95.

- **Groupe de travail**

Lors de la réunion de lancement du projet, un groupe de travail a été constitué afin de mener à bien cette mission, auquel a participé une partie de l'équipe OTC :

⇒ la responsable du portail Cyberthèque
⇒ la responsable publications
⇒ le responsable études stratégiques

Il a aussi été décidé de s'adresser à des « experts » internes Société Générale pour des précisions supplémentaires concernant les domaines très techniques (ex. : sécurité, architecture).

- **Suivi du projet**

⇒ Réunion hebdomadaire

Lors de la réunion de lancement de projet tenue le 7 juillet 2003, il a été convenu de la tenue d'une réunion hebdomadaire du groupe de travail, d'une durée minimale d'une demi-heure, afin d'assurer la continuité et le suivi du projet et de rendre compte de son état d'avancement.
Le compte-rendu de chaque réunion a été distribué à chaque membre du groupe de travail.

⇒ ... mise à jour des documents de travail

Après chaque séance de travail, les modifications proposées ont été reportées sur les documents de travail et un nouveau document édité et distribué aux membres du groupe de travail, comme base pour la séance suivante.

2.1.2.2 CONCLUSIONS DE L'AUDIT

L'audit a été exposé et ses conclusions livrées lors de la réunion de service de INFO/STA/OTC, le 11 août 2003, à l'aide d'une présentation Power Point. La liste des réalisations possibles dans le temps imparti et des réalisations possibles dans un laps de temps supplémentaire a été exposée avant de passer à la phase de réalisation des modifications dans l'interface de gestion des concepts Verity Intelligent Classifier.

2.1.2.3 CHOIX DE LA MÉTHODE : DÉMARCHE DE MISE À JOUR DU LANGAGE CONTRÔLÉ

Étant donnée la spécificité de la mission, du service et du fonds géré, l'option a été prise d'effectuer une mise à jour partielle et d'utiliser une méthode de travail fortement adaptée au service et au profil des veilleurs.

La méthode de mise à jour de langage contrôlé adoptée est la méthode « bottom up ». La collecte des nouveaux termes a été faite principalement lors des séances de travail avec les veilleurs faisant partie du groupe de travail et lors des entretiens avec les experts, suivie de la vérification des termes recueillis à l'aide des dictionnaires informatiques disponibles en ligne. L'établissement des synonymies manquantes a été fait entre les séances de travail et vérifié lors des séances de travail et des entretiens. La tentative de validation des hiérarchies en place, lors des entretiens avec les experts, n'a pas abouti.

Les experts n'ayant pas validé la hiérarchisation de certains termes, car les notions de même degré d'abstraction n'étaient pas regroupées aux mêmes niveaux, ni dans les concepts, ni dans la taxonomie, la mise à jour s'est arrêtée à ce stade. En effet, le temps de réalisation de la mission ne permettait pas d'effectuer la mise à jour des relations hiérarchiques de chaque concept, compte tenu de la fréquence des cas de polyhiérarchie.

L'implication de l'équipe dans le projet était une des conditions nécessaires à l'avancement de la mission. Le groupe de travail, créé dès le début, a bien fonctionné, malgré les séances à effectif réduit pendant les congés de chacun. La démarche choisie a ainsi permis de créer une dynamique autour du travail effectué et d'impliquer l'équipe dans le projet, ce qui ne paraissait pas évident au début. Les entretiens avec les experts du domaine afin de valider les concepts et leur structuration, ont permis de mieux appréhender le type de travail à effectuer.

Le choix de sensibiliser l'équipe au type de travail que cela représente a été fait dans le but de mettre les bases d'une future mise à jour.

2.1.2.4 PHASE OPÉRATIONNELLE

La phase d'évaluation a été suivie d'une phase opérationnelle qui a duré 4 semaines (18 août - 12 septembre 2003), pendant laquelle une partie des modifications et mises à jour proposées ont été mises en œuvre dans l'application Verity Intelligent Classifier, tout en continuant les séances de travail avec le groupe et les entretiens avec les experts.

2.1.2.5 FIN DU PROJET

La réunion de fin de projet a eu lieu le 11 septembre 2003. Lors de cette réunion, les membres du groupe de travail ont pris connaissance des conclusions concernant la mise en place des changements et la fin du projet a été validée.

2.2 AUDIT DES CONCEPTS ET DE LA TAXONOMIE

2.2.1 ANALYSE DES BESOINS

2.2.1.1 PROBLÈMES ÉVOQUÉS

Les problèmes rencontrés par l'équipe OTC dans sa pratique quotidienne ont été clairement évoqués dès février 2003 :

- **Pertinence de certains résultats d'interrogation**

La pertinence des documents retournés suite aux interrogations ne semble pas satisfaisante à l'équipe OTC.

- **Bruit**

Avec l'augmentation du nombre de documents dans la base, les documents pertinents semblent de plus en plus difficiles à retrouver, noyés dans une masse de réponses de plus en plus longue.

- **Taxonomie à revoir**

L'équipe OTC a exprimé l'impression que le moteur se comportait de façon incompréhensible, notamment concernant l'attachement des documents aux classes préétablies : « ça classe tout dans tout ».

- **Concepts et noms de produits nouveaux à ajouter**

Le domaine de l'informatique et des technologies appliquées au domaine bancaire est en plein mouvement. Des nouvelles pratiques apparaissent, liées à Internet et à la vente de services en ligne. Une partie des mots recensés sont des anglicismes. Par exemple, on utilise « e-payment » pour les paiements par Internet, « m-commerce » pour « mobile-commerce », etc.

- **Gestion des équivalences français-anglais**

Étant donné la présence de documents en anglais en plus des documents en français, il est nécessaire de les retrouver. Le problème du multilinguisme n'a pas pu être géré par la version du moteur installée (4.0.1). Selon l'équipe commerciale de l'éditeur, le multilinguisme (cross-language) peut être géré par la version supérieure (version 5, commercialisée en juin 2003). En attendant, les équivalences sont établies via la gestion des « concepts ».

- **Étonnement concernant quelques « concepts »**

Par exemple, « SAN » (Storage Area Network) avait donné comme résultats d'interrogation « sans » et donne toujours : « San Francisco », « San José ». Ces résultats, qui n'ont aucune pertinence, se retrouvent quelquefois classés très haut dans la liste des réponses. L'équipe OTC souhaitait en comprendre les raisons.

2.2.1.2 OBJECTIFS

Au regard des problèmes évoqués, les objectifs à atteindre sont de plusieurs niveaux : concepts, taxonomie, méthode de travail de mise à jour du dictionnaire métier.

- **Concepts**

Concernant les concepts, sur lesquels le moteur s'appuie lors de la recherche simple et de la recherche avancée, les objectifs à atteindre sont :

⇒ Augmenter la pertinence des résultats d'interrogation
⇒ Réduire le bruit
⇒ Meilleur classement des documents selon les poids

- **Taxonomie**

Les changements effectués à ce niveau sont directement visibles dans le plan de classement du portail, lorsqu'on choisit d'effectuer une recherche par « Navigation thématique ». Le résultat attendu serait donc :

⇒ Meilleure répartition (et représentation) des connaissances

- **Méthode de travail**

Jusqu'ici, la mise à jour des concepts et de la taxonomie a été faite à la volée, selon des besoins ponctuels issus de la veille. L'absence de documents listant l'existant rend cette opération aléatoire et le risque de créer des doublons est relativement important.

D'autre part, lors d'une mise à jour globale, pour se faire une idée de l'existant, il s'avère nécessaire d'exporter le fichier en format propriétaire Verity, le transformer en fichier texte (.txt) et le retraiter ensuite sous Word. Les objectifs fixés sont donc :

⇒ Augmentation du confort de la mise à jour
⇒ Gestion du dictionnaire métier basée sur des critères formalisés et logiques

2.2.2 ANALYSE DE L'EXISTANT

2.2.2.1 CONTRAINTES

2.2.2.1.1 MATÉRIELLES

Deux contraintes principales peuvent être rappelées :

- **Temps de réalisation de la mission (2 mois)**

La première contrainte d'ordre matériel est le temps, la mission de stage s'effectuant en deux mois. Le laps de temps accordé à la mise à jour d'un langage contrôlé généralement est largement supérieur.

- **Absence de certains acteurs (période de congés)**

La seconde contrainte est liée à la présence des acteurs impliqués dans le projet pendant la période de la mission. En effet, les mois de juillet et août sont des périodes de congés annuels, donc tous les acteurs s'absentent pendant des intervalles de temps de deux à trois semaines.

Ces contraintes ont été prises en compte et des efforts ont été faits pour pallier l'absence de certains acteurs à des moments clé, notamment en maintenant le dialogue et en tenant à jour les outils de suivi du projet.

2.2.2.1.2 OPÉRATIONNELLES

Le même set de concepts est utilisé pour gérer l'information structurée (bases Lotus Notes) et de l'information non-structurée (pages Web indexées par le spider).

En regardant de plus près les champs indexés des bases Notes, le moteur Verity K2 indexe les champs suivants (leurs contrôles sont rappelés entre parenthèses) :

⇒ Titre (texte libre)
⇒ Résumé (texte libre)
⇒ Date (format de date : jj/mm/aaaa)

Parmi les champs non indexés, on peut citer le champ « Source ».

En conclusion, bien que stockée dans une base de données, il s'agit d'information non-structurée : que ce soit le titre, le résumé ou le document lié.

Ceci a une incidence sur l'usage des opérateurs, car une partie des opérateurs sont à utiliser uniquement sur champs (<Field> = <Champ>). On peut quand même se servir de l'opérateur sur zone (<In> = <Dans>), pour peu que la zone ait été définie lors de la création de la collection, or toutes les zones n'ont pas été définies.

2.2.2.2 ÉVALUATION DES CONCEPTS ET DE LA TAXONOMIE

Les 345 « concepts » (Topics) présents dans l'application ont été étudiés et testés en interrogation à travers l'interface Intelligent Classifier sur des bases de test contenant 3437 articles sur les approximativement 4800 actuellement présents dans le portail. Les points positifs et les points négatifs constatés sont présentés succinctement :

2.2.2.2.1 POINTS POSITIFS

Les points positifs relevés pendant l'analyse de l'existant sont les suivants :

- **Concepts et taxonomie**

Concernant les noms des concepts (référentiel ou dictionnaire métier), le constat est globalement positif :

⇒ Termes pertinents

La pertinence des termes choisis n'est pas à démontrer, ce sont les termes de travail de tous les jours des veilleurs.

⇒ Couvrent les domaines qui intéressent OTC

Les domaines d'intérêt (Banque et Informatique), sont globalement bien couverts, mais certaines notions manquent et des notions nouvelles ne sont pas encore intégrées.

- **Hiérarchie**

Les cas de polyhiérarchie sont très nombreux dans la base de connaissances. De ce point de vue, on peut apprécier :

⇒ Bonne résolution de certains problèmes de polyhiérarchie

Certains cas de polyhiérarchie sont bien résolus, comme, par exemple, « chiffres clés ». Ce concept est classé dans chaque catégorie de la taxonomie, mais sa règle est définie en utilisant l'opérateur booléen non-flou <And> et donne des résultats corrects en interrogation.

- **Opérateurs**

Concernant les opérateurs (très peu nombreux) qui ont été utilisés, on peut relever un point positif :

⇒ Syntaxe correcte

La syntaxe est validée par le parser IC lors de la création des concepts, il est donc impossible de créer des concepts à syntaxe invalide.

2.2.2.2.2 POINTS NÉGATIFS

Les points négatifs relevés lors de l'analyse des concepts et de la taxonomie sont listés par catégories : problèmes rencontrés dans la définition des concepts, dans les règles d'inférence (« business rules » ou « classify rules ») de la taxonomie, gestion de la hiérarchie et utilisation des opérateurs.

- **Concepts**

Plusieurs types de problèmes ont été constatés concernant la gestion des concepts :

⇒ Concepts trop touffus

Certains concepts (topics) sont tellement touffus qu'ils remontent à l'interrogation le tiers de la base. Parmi les concepts sous lesquels trop de choses ont été rangées, on peut citer « concurrence bancaire », qui remonte à l'interrogation 1499 documents sur les 3437 (collections de test).

⇒ Concepts orphelins

Certains concepts ne sont reliés à aucun autre concept et à aucune catégorie dans la taxonomie : ils semblent avoir été rajoutés sans vérification préalable. Ils sont donc têtes de hiérarchie dans le Topic Set, mais ne contiennent rien d'autre et ne font pas partie de la taxonomie (par exemple, JPEG, MPEG, API).

⇒ Doublons

Certains sont des doublons de concepts contenant une faute d'orthographe ou une erreur dans la spécification de l'instanciation (Evidence Level).

C'est le cas de MPEG, défini comme « Motion Picture Expert Group ». Or, il s'agit d'une norme traitant l'image en mouvement par opposition à l'image fixe, non pas l'industrie cinématographique. Le nom exact de la norme est « Moving Picture Experts Group ».

Un autre exemple est ASP, défini deux fois dans les concepts. La première fois, il y a uniquement le sigle, au niveau instanciation d'une hiérarchie. La seconde fois, ASP est tête de hiérarchie et est défini uniquement comme « Application Server Provider », notion qui n'apparaît jamais dans les textes contenus dans la base. Cette notion est considérée dans les dictionnaires de terminologie informatique comme un quasi-synonyme de « Application Service Provider », qui figure dans les documents. « Application Server Provider » est une partie du service offert en mode ASP.

En interrogation, les deux exemples donnent la réponse « 0 résultats », malgré la présence de documents dans la base.

❑ Taxonomie

L'organisation des niveaux de hiérarchie de la taxonomie n'a pas été validée par les spécialistes du domaine consultés[28].

⇒ Règles d'inférence (Classify Rules) mal écrites

Les règles d'inférence qui font qu'un document est classé dans telle ou telle classe ne semblent pas obéir à des formalisations précises.

❑ Hiérarchie

Autant dans le cas des concepts que dans celui de la taxonomie, la hiérarchie est conçue de façon assez surprenante. Il est globalement difficile de s'y retrouver lors de la mise à jour.

⇒ Polyhiérarchie globalement mal gérée

La polyhiérarchie est un des problèmes de cette base, car elle ne semble pas avoir été gérée selon des critères logiques.

La polyhiérarchie est globalement mal gérée dans la taxonomie, à part le cas cité dans les points positifs.

Concernant les concepts et leur lien avec la taxonomie, la gestion actuelle de la polyhiérarchie pose des problèmes de classement. Des concepts comme « acteurs », « opérateurs télécoms », « opérateurs mobile », très riches, se retrouvent dans toutes les catégories. Pour cette raison, effectivement, on peut dire, comme l'avait constaté l'équipe, que « ça classe tout dans tout ».

⇒ Noeuds de même niveau / information hétérogène

On retrouve au même niveau de l'information hétérogène qui serait à classer à des niveaux de profondeur différents, car les concepts expriment des notions de niveau d'abstraction (ou de généralisation) différent.

⇒ Mélange produits / services / protocoles

On retrouve mélangées au même niveau des notions qui devraient être regroupées en catégories distinctes selon des critères formalisés, comme, par exemple : produits d'un côté, services de l'autre, protocoles etc.

⇒ Ordre des noeuds à l'affichage : aucun

Les concepts têtes de hiérarchie sont rangés par ordre alphabétique par l'application lors de leur création. Mais lorsqu'un concept est déployé, qu'il s'agisse du deuxième ou du n^{ème} niveau, aucun ordre n'a été respecté.

⇒ Mélange entre « subtopics » et « evidence level »

Quelquefois le niveau « Evidence level » n'est pas renseigné avec le nom du subtopic ou du concept, pourtant il est obligatoire et indispensable : c'est le seul lien entre le texte des documents et les concepts : le lien par lequel le Topic Set pointe vers les documents.

[28] L'équipe OTC avait décidé d'adopter le point de vue métier (banquier, juriste) et non le point de vue expert technique (technologies de l'information) lors du choix des points d'entrée, début 2002.

❑ Opérateurs

Comparée à la richesse du langage de requête de Verity (VQL) et à ses possibilités, la pauvreté de la syntaxe utilisée dans l'application est une des causes principales des problèmes constatés.

⇒ Peu d'opérateurs utilisés : 7 sur les plus de 35 possibles

Pour définir les concepts, 5 opérateurs ont été utilisés de façon systématique, sans distinction (<Cumul>, <Ou>, <Plusieurs>, <Expression>, <Racine>), et 2 autres apparaissent de façon très sporadique (<Mot>, <Proche>).

⇒ Opérateurs inadaptés

Parmi le peu d'opérateurs utilisés, certains ont été utilisés de façon très incorrecte. C'est le cas de l'opérateur <Racine> (<Stem>). Cet opérateur demande la lemmatisation du concept employé. Or, il a été utilisé dans la définition de tous les sigles et de tous les noms propres, et le vocabulaire concerné en compte bon nombre. L'opérateur <Casse>, dont l'usage aurait pu résoudre quelques problèmes de bruit concernant les sigles, n'a jamais été utilisé.

La lemmatisation est une opération valide uniquement pour les catégories grammaticales flexibles (= ont un pluriel, un féminin ; ex. ; noms, verbes). Les sigles n'ont pas de « racine », ni de catégories grammaticales à lemmatiser, donc, l'opérateur <Racine> est inadapté au résultat souhaité. Cet usage incorrect de l'opérateur racine explique les problèmes relevés par l'équipe concernant les résultats d'interrogation du sigle « SAN » (= Storage Area Network), qui donnait des résultats pour le moins surprenants, comme « sans », « San Francisco » etc.

L'opérateur <Racine> a aussi été utilisé pour demander la lemmatisation de tous mots anglais, mais le lemmatiseur anglais n'est pas présent dans l'application, car la « locale » anglais n'a pas été installée. Rappelons que le moteur Verity est un système qui fonctionne avec un dictionnaire de lemmes et que le dictionnaire est fourni par l'installation des « locales ». Par conséquent, on ne peut pas lemmatiser des mots absents du dictionnaire, à plus forte raison si le dictionnaire n'est pas installé, car les lemmes anglais ne figurent pas dans le dictionnaire français.

Les noms propres n'ont pas de racine non plus, car ils représentent des objets ou des personnes ayant un référent unique dans le monde réel. Les noms propres ne se déclinent pas en français (on n'écrit pas « *les Dupont**s** », mais « les Dupont »), donc ils n'ont pas de catégories à lemmatiser. Par conséquent, pour peu qu'un nom propre s'écrive de la même façon qu'une forme verbale, en l'absence de l'opérateur <Casse>, qui peut être utilisé pour bloquer la recherche uniquement sur les mots écrits en majuscule (ou avec une première lettre en majuscule), il sera ramené au lemme présent dans le dictionnaire de lemmes. Par exemple, « SEMA », personne morale, défini à l'aide de l'opérateur <Racine>, a donné des résultats d'interrogation « Il n'y a pas de quoi **semer** la panique. », à cause de la forme verbale « sema » (3^{e} personne du passé simple de l'indicatif).

2.2.2.2.3 CHIFFRES

Le nombre réduit de « concepts » (« Topics ») présents dans l'application, 345, est à mettre en relation avec le nombre de notions non définies au niveau des instanciations (« evidence level »).

Si on prend en compte toutes les notions (termes) recensées (775) et les termes synonymes 182), on peut tirer la conclusion qu'il y a 957 notions au total (définis comme « Topics » ou pas).

La Figure 22 - Notions et concepts (quelques chiffres), ci-dessous, extraite du document Power Point utilisé lors de la réunion de service du 11 août 2003 pour communiquer les conclusions de l'audit, présente ces chiffres de façon synthétique.

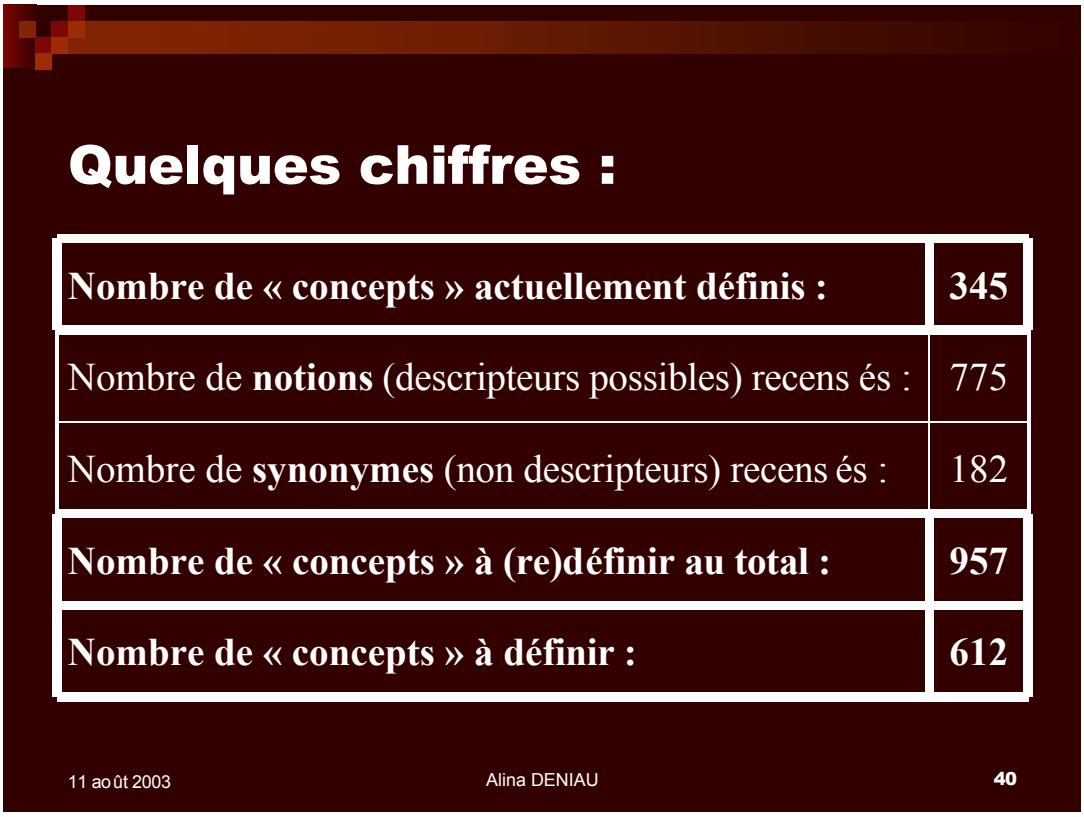


Figure 22 - Notions et concepts (quelques chiffres)

2.2.3 RÉSULTATS DE L'AUDIT

2.2.3.1 CONCLUSIONS DE L'AUDIT

Les conclusions de l'audit, présentées à la réunion de service du 12 août 2003 font état des points suivants, nécessaires à l'amélioration de ce qui a été mis en place :

- Vérifier les requêtes de chaque concept (345)
- Réécrire les requêtes des sigles
- Ajouter les notions nouvelles
- Définir les 612 concepts manquants
- Éliminer les doublons
- Restreindre (voire éliminer) les cas de polyhiérarchie
- Réécrire les règles d'inférence de la taxonomie
- Éditer un document sur l'existant et les opérateurs

Il était évident que le laps de temps d'un mois (entre le 12 août et le 12 septembre 2003) était insuffisant pour mettre en place tous ces changements et valider avec les spécialistes du domaine les éventuels changements dans la taxonomie, qui entraînent des modifications visibles dans le portail en navigation thématique.

En effet, la mise en place soit d'une taxonomie consensuelle reflétant les différents points de vue, soit de plusieurs taxonomies métier, était un travail de longue haleine, qui ne pouvait pas être effectué dans le cadre de la mission de stage.

2.2.3.2 CONSEILS POUR AMÉLIORER LA GESTION DES CONCEPTS ET DE LA TAXONOMIE

❑ Principes de travail

Quelques principes de travail peuvent être rappelés :

⇒ Toujours garder à l'esprit le fait qu'il s'agit d'abord d'une construction intellectuelle (concepts), cohérente

⇒ Chercher le moyen de l'exprimer en utilisant les possibilités de l'outil Verity Intelligent Classifier

⇒ Être conscient des limites du logiciel :
- il ne « comprend » pas le langage humain
- il ne « réfléchit » pas à la pertinence de l'indexation
- il fait seulement ce qu'on lui demande de faire

⇒ Être conscient des avantages du logiciel :
- les temps de traitements courts
- la recherche par filtrage conceptuel est très performante si les concepts ont été bien définis
- la recherche full-text est très performante (pour les notions nouvelles)

❑ Méthode de travail

Quelques conseils d'ordre concret peuvent servir de guide pour la gestion des concepts avec Intelligent Classifier :

⇒ Classement alphabétique
- des sous-niveaux (« subtopics », en bleu à l'affichage)
- des instanciations (« evidence level », en gris à l'affichage)

⇒ Nœuds de même niveau = notions de même niveau

⇒ Synonymes : à ranger juste après le nom du concept, opérateur <Ou>

⇒ Choix des opérateurs appropriés dans l'écriture des requêtes

⇒ Mise à jour des concepts : d'abord sur papier pour garantir la cohérence

⇒ Garder une trace : document interne daté (fichier et papier) de l'existant à un moment T, le mettre à jour

⇒ Recommandations concernant les sigles :
- Ne pas utiliser l'opérateur <Racine> pour les sigles
- Utiliser l'opérateur <Casse> pour les sigles au niveau « evidence »
- Exclure, éventuellement, les sources de bruit recensées en utilisant l'opérateur <Sauf> (mais on risque d'exclure des documents pertinents)

2.2.3.3 QUELQUES RÉALISATIONS

Les réalisations jugées possibles entre le 12 août et le 12 septembre 2003 ont été mises en place dans le module de gestion Intelligent Classifier pendant la phase opérationnelle :

- Réécrire les requêtes des sigles
- Éliminer une partie des doublons
- Ajouter des notions nouvelles
- Éditer un document sur les opérateurs, l'existant

Le document listant les opérateurs et leur comportement est présenté en annexe, page 150. Il a été remis aux membres du groupe de travail.

Une sortie imprimante du document listant l'existant et les modifications apportées a été distribué à chaque membre du groupe de travail lors de la réunion de fin de projet, le 11 septembre 2003.

Les deux fichiers, nommés de façon parlante, sont placés dans un répertoire sur le disque partagé réservé au service et accessible à toute l'équipe.

Les doublons ont été éliminés et les concepts orphelins replacés dans leur hiérarchie.

Une partie des requêtes des sigles, des noms propres et des mots anglais ont été réécrites en utilisant les opérateurs adéquats : <Mot>, <Casse>, selon le cas. Les changements ont été validés en testant les résultats d'interrogation dans Intelligent Classifier par la visualisation des documents (surbrillance des mots de la requête et de ceux apportés par l'expansion de requête à travers les concepts).

La mise en place des changements concernant les opérateurs a donné de bons résultats en test. La pertinence des résultats d'interrogation a augmenté, tout en faisant baisser le bruit apporté par la lemmatisation intempestive des sigles et des noms propres.

2.3 RECOMMANDATIONS

Au-delà des quelques conseils formulés précédemment, il est nécessaire de considérer la gestion à moyen et long terme. Plusieurs scenarii sont envisageables afin d'optimiser la gestion des concepts et de la taxonomie.

2.3.1 SCENARII POSSIBLES

2.3.1.1 MISE À JOUR AU FUR ET À MESURE

Ce choix nécessite un investissement supplémentaire de la part de l'équipe OTC. Or, les veilleurs ont déjà de nombreux objectifs à atteindre dans leur pratique professionnelle courante (édition des revues de presse mensuelles, synthèses, maîtrise d'ouvrage des développements informatiques, relations avec les cabinets d'analystes, etc.)

Toutefois, dans cette optique, un tableau du fonctionnement des opérateurs de Verity avec quelques conseils d'utilisation est présenté en annexe, page 150. Cela peut permettre une gestion des petits rajouts sporadiques. Mais pour une gestion efficace, il est recommandé de se reporter aussi souvent que possible à la documentation de Verity, notamment en cas de doute. Parmi les guides fournis par le constructeur en format papier, trois documents sont indispensables pour gérer les mises à jour :

- ⇒ Intelligent Classification Guide, (Verity, 2001d, [80])
- ⇒ Verity Query Language Guide (Verity, 2001f, [83])
- ⇒ Fundamentals Guide (Verity, 2001c, [73])

Les avantages et inconvénients de ce mode de gestion sont :

- **Avantage**

Ce mode de gestion ne semble entraîner aucun coût à première vue, car la mise à jour est effectuée par le personnel déjà en place, sur son temps de travail. Mais la mise à jour du langage contrôlé n'est pas l'activité première d'aucun des membres de l'équipe OTC, donc, les rajouts effectués ont un caractère peu formalisé.

- **Inconvénient**

Le risque majeur de ce mode de gestion est la perte de cohérence de la base de connaissances dans le cas où les seules mises à jour seraient celles effectuées au fil de l'eau. À terme, ce mode de gestion peut entraîner des coûts très importants, car une mise à plat s'avèrera vite indispensable, car la qualité d'une base de connaissances se dégrade très vite dans ces conditions, et ne donnera plus aucune satisfaction en interrogation.

2.3.1.2 MISE À JOUR PÉRIODIQUE

Compte tenu de l'évolution rapide du domaine traité (informatique et technologies), une mise à jour annuelle serait souhaitable. Un délai plus grand n'est pas envisageable, car, dans ce cas, la pertinence des résultats d'interrogation baissera rapidement et une mise à plat complète deviendra nécessaire, ce qui entraînera inévitablement des coûts encore plus importants.

Un délai de deux ans entre deux mises à jour paraît être un seuil critique au-delà duquel la mise à plat s'avèrera indispensable.

Cette mise à jour annuelle devra être effectuée par un professionnel de la gestion des langages contrôlés, idéalement par quelqu'un qui maîtrise déjà le langage de Vérity. Il serait conseillé d'éviter la période des congés d'été, pour que tous les acteurs puissent être présents.

Un consultant en gestion des langages documentaires (thésaurus, taxonomies) serait le profil adapté. Ils sont à chercher du côté des cabinets conseil en ingénierie documentaire et des consultants indépendants.

- **Avantages**

L'avantage majeur de ce mode de gestion est la grande qualité des résultats, due aux connaissances et à l'expérience du consultant amené à intervenir.

D'autre part, ce mode de gestion ne nécessite pas d'investissement en termes de formation du personnel, qui peut néanmoins bénéficier d'une expérience en la matière en participant au groupe de travail.

Un autre avantage, non-négligeable lui non plus, est le regard extérieur, qui permet à l'équipe de faire le point sur son travail et de se poser des questions qui surgissent uniquement lorsqu'il s'agit de projets.

- **Inconvénient**

Le seul inconvénient de ce type de gestion peut être le prix : l'emploi d'un consultant peut revenir cher.

Mais ce désavantage est très relatif en comparaison avec les dépenses que peut entraîner la gestion au fil de l'eau.

2.3.1.3 MISE À JOUR PAR DES STAGIAIRES

Si, pour des raisons d'ordre financier, le service ne souhaite pas faire appel à un consultant, une solution de compromis consisterait à confier cette mission à des stagiaires en formation en DESS information-documentation (I&D).

Quelques conseils utiles pour le choix d'un stagiaire :

⇒ choisir un stagiaire qui s'intéresse à l'exercice du thésaurus
Dans certaines écoles, l'exercice sera en cours, dans d'autres il sera à faire assez tard dans l'année et ne pourra pas servir de base au stagiaire, n'ayant pas le recul nécessaire pour aborder la nouvelle expérience. Poser des questions à ce sujet.

⇒ de préférence ayant une formation de base de linguiste
C'est un profil rare dans la profession. Penser à le spécifier en rédigeant la fiche ou l'offre de stage.

⇒ ayant des qualités relationnelles
Toute négociation qui touche au langage prend du temps et demande des qualités d'écoute, une certaine diplomatie, notamment pour gérer le groupe de travail et surtout les rencontres avec les spécialistes.

- **Avantages**

L'avantage majeur de cette « solution » est, évidemment, le prix.

Un autre avantage est le côté formateur pour l'étudiant.

- **Inconvénients**

Le premier inconvénient est le temps : sur les 3 mois de stage, 2 mois peuvent être consacrés à la réalisation de la mission. C'est insuffisant pour une personne ayant à se former seule à la gestion de l'interface à l'aide de la documentation éditeur (en anglais), et à devoir gérer seule ce que les consultants estiment à plusieurs mois-hommes de travail effectif.

Le second désavantage peut être le manque d'expérience du stagiaire, auquel peut s'ajouter le manque de maturité. Or, il s'agit d'un travail sensible, qui peut bousculer les habitudes : des qualités relationnelles sont nécessaires pour mener à bien la mission de stage... Dans le cas d'un stagiaire très jeune, il peut se trouver quelque peu déboussolé.

On peut citer aussi la nécessité d'encadrement, car le service ne semble pas avoir mis en place de procédures spécifiques pour accueillir les stagiaires. L'étudiant peut se retrouver en électron libre s'il manque d'initiative et d'assurance.

2.3.2 CHOIX RECOMMANDÉ

Compte tenu des conclusions de l'audit, la seconde solution, qui consiste à faire appel à un consultant spécialisé, est fortement recommandable. Ses avantages sont indéniables comparés au seul désavantage : le prix. Mais le service (et, plus généralement, l'entreprise) est habitué à faire appel à la sous-traitance par des spécialistes du domaine, même dans des domaines sensibles comme, par exemple, la sécurité informatique.

Faire appel, de façon périodique, à un spécialiste de la gestion de taxonomies, thésaurus et référentiels d'entreprise est la solution la plus adaptée à la spécificité du service et du fonds géré et à la complexité de gestion des concepts à l'aide de l'interface Verity Intelligent Classifier.

La formation à l'utilisation de l'interface de gestion Intelligent Classifier, dispensée par l'éditeur Verity, ne saurait remplacer une formation théorique et pratique à la gestion de vocabulaire, ni l'expérience acquise dans la construction et l'utilisation de vocabulaires contrôlés (comme les thésaurus, listes d'autorité, classifications décimales). Il s'agit de techniques et de méthodes de travail intellectuel qui se conçoivent indépendamment de tout outil informatique mais s'appliquent à tous ces outils, moyennant l'adaptation aux spécificités de l'interface, du service ou de la branche d'activité.

Le travail de mise à jour d'un langage contrôlé qui n'est pas géré par un spécialiste de la chose tout au long de l'année est un travail pointu, à confier à des consultants spécialisés, car il fait appel à des techniques précises, mais aussi à une grande capacité d'adaptation aux structures et aux codes de l'entreprise.

En outre, la présence périodique d'un consultant menant un projet peut avoir des effets bénéfiques sur le travail de l'équipe.

CONCLUSION

Les outils existent. Ils sont complexes et paraissent mystérieux à première vue. Cette impression tient autant à leur complexité qu'à la stratégie commerciale. D'autre part, pourrait-on imaginer un salon qui prend des airs d'amphithéâtre, où l'on vous expliquerait, plusieurs heures durant, le pourquoi du comment afin de vous vendre un logiciel ?

Peut-être que la solution au problème consiste à s'entourer de professionnels, qui connaissent les produits et comprennent les besoins, afin d'orienter le choix et intervenir dans la mise en place des applications.

❑ Gestion de langage documentaire (langage contrôlé)

Une autre question qui se pose est :
« Est-ce que la gestion de taxonomies, ontologies, Topic Maps est une forme nouvelle de gestion de langage documentaire ? »

Après avoir fait le point sur les outils informatiques et les outils « intellectuels », répondre par l'affirmative devient possible, et ce pour plusieurs raisons :

⇒ les techniques de construction de thésaurus participent à la modélisation du savoir afin de rendre le modèle interprétable par la machine

⇒ ces techniques, formalisées depuis près de 30 ans (la première version de la norme ISO 2788 date de 1974) et appliquées avec succès, peuvent s'adapter aux outils existants moyennant un appauvrissement des relations gérées (taxonomies) ou, bien au contraire, leur enrichissement (ontologies, Topic Maps)

⇒ le contrôle du langage utilisé par les moteurs de recherche linguistiques nécessite la conjugaison de la maîtrise des techniques de construction de thésaurus et de la connaissance du fonctionnement des différents types d'opérateurs, pas uniquement booléens (voir, à ce sujet, l'exemple de Verity)

Selon une étude récente de Giga Information Group, de la série Planning Assumption (Ramos, 2003, p. 6, [21]), la gestion de taxonomies nécessite des compétences documentaires (en anglais : « library scientist »), qu'elle soit construite en mode manuel ou générée en mode automatique :

> *« [...] library scientist or people with linguistics background are required to develop and maintain the taxonomy [...] »*

Le même article estime que l'entretien d'une taxonomie, sa mise à jour, est une opération plus coûteuse que sa mise en place. Elle est néanmoins indispensable pour éviter le danger de posséder une structure rigide qui n'évolue pas au même rythme que les changements qui s'opèrent dans l'entreprise (Ramos, 2003, p. 6, [21]).

L'étude IDC de 2001, « The High Cost of Not Finding Information » (IDC, 2001, p. 9, [3]) estime qu'une gestion défectueuse de la circulation de l'information dans l'entreprise peut lui faire manquer des opportunités faute d'avoir possédé la bonne information au moment de prendre des décisions. L'étude va jusqu'à chiffrer les pertes, fait valoir que la vie de l'entreprise dépend de ses décisions stratégiques, et conclut sur la nécessité de faire appel aux meilleurs outils afin de rendre accessible l'information :

> *« Decisions are usually information problems. If they are made with poor or erroneous information, then they put the life of the enterprise at stake. Therefore, it behooves the entreprise to provide the best information-finding tools available and to ensure that all its intellectual assets have access to them, no matter where they reside. »*

Parmi les meilleurs outils (« best information-findings tools ») à mettre en place pour faciliter la recherche d'information sur les intranets des entreprises, les moteurs de recherche s'avèrent indispensables. Et, pour construire les meilleurs référentiels d'entreprise, taxonomies et thésaurus destinés à garantir la qualité des résultats d'interrogation obtenus par les moteurs de recherche, il convient de faire appel aux meilleurs professionnels.

❑ Métier de documentaliste : nécessité de s'adapter

Pour le métier de documentaliste, un métier en perpétuel changement, qui depuis longtemps suit les évolutions technologiques, la nécessité d'adaptation fait partie des constantes. Il s'agit de s'adapter à nouveau :

⇒ à ce nouveau modèle : moteurs de recherche avec interfaces fortement inspirées du web, différents des logiciels documentaires et des bases de données en ligne

⇒ au travail hors centre de documentation, dans des services qui gèrent uniquement de la documentation électronique (pas de livres, pas de revues)

⇒ au travail avec des équipes n'ayant aucune notion de gestion de langage documentaire

Les enjeux de la recherche d'information en entreprise sont importants. Les grands comptes, qui ont les moyens de mettre en place des outils coûteux, mais néanmoins indispensables, comme les moteurs de recherche, ont besoin actuellement de pouvoir assurer la gestion des référentiels d'entreprise.

Les personnes à même de le faire dans les meilleures conditions sont les ingénieurs documentalistes, professionnels de l'information-documentation (I&D) possédant les connaissances nécessaires pour prendre en charge ce type de missions, en comprendre les enjeux et proposer des solutions adaptées aux services ayant en charge les outils.

Chaque réussite valorise notre métier et fait prendre conscience aux gens qu'il est intéressant et qu'il nécessite un haut degré de connaissances et de technicité.

BIBLIOGRAPHIE

PRÉSENTATION

La bibliographie a été arrêtée en octobre 2003.

Les notices sont rédigées selon les normes bibliographiques suivantes :

⇒ NF Z 44-005 (décembre 1987, confirmée 1992) pour les monographies et les articles
⇒ NF ISO 690-2 (février 1998) pour les ressources électroniques

Deux présentations sont proposées dans le souci de faciliter une éventuelle recherche bibliographique :

- **Bibliographie analytique**

Numérotée et classée par ordre d'apparition des renvois dans le texte, la bibliographie analytique est présentée en premier. Le plan de classement des notices suit le plan du mémoire et celles-ce sont accompagnées d'un résumé. La plupart des résumés des articles proviennent de la Banque de Données BD-ID, constituée par le Centre de Ressources Documentaires (CRD) de l'INTD (dont est extrait le Bulletin Bibliographique de l'INTD). Ils ont été conservés en l'état dans la majorité des cas. Les résumés des monographies et des ressources électroniques ont été rédigés pour l'occasion. Les notices des monographies proviennent, en partie, du catalogue Opale Plus de la bibliothèque nationale de France (BnF).

- **Bibliographie alphabétique des noms d'auteurs**

Une deuxième présentation, par ordre alphabétique des noms d'auteurs (personnes physiques ou morales) accompagnés de la date de parution de l'ouvrage, permet de retrouver plus facilement la référence d'un article ou d'un ouvrage potentiellement intéressant. Pour les actes de congrès, le nom du directeur de publication ou du premier éditeur intellectuel a été choisi comme clé. Pour les ouvrages collectifs (comme le Dictionnaire encyclopédique de l'information et de la documentation, Nathan, 1997), le nom du premier auteur a été (arbitrairement) retenu comme clé. Les notices ne sont pas accompagnées de résumés.

Afin de faciliter l'identification des documents, quelques icônes ont été utilisées :

🕮	=	monographie
🗎	=	article
🗏	=	article extrait d'une base de données payante
💻	=	ressource électronique
✍	=	mémoire INTD

BIBLIOGRAPHIE ANALYTIQUE

1 MOTEURS DE RECHERCHE

1.1 LES MOTEURS DE RECHERCHE ET LA RECHERCHE D'INFORMATION

[1] Maniez, 1999

📖 **Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information** / [textes réunis par] Jacques Maniez, Widad Mustafa El Hadi. — Villeneuve d'Ascq : Université de Lille 3, 1999. — 403 p. — (Travaux et recherches — Université de Lille 3). — Notes bibliogr. Index. — ISBN 2-84467-002-4

Rassemble des études de portée générale concernant les pratiques d'indexation et des approfondissements plus ciblés sur l'amélioration de la recherche d'information, notamment par le perfectionnement des techniques d'indexation automatique.

[2] Le Moal, 2002

📖 **La recherche d'information sur les réseaux : cours INRIA, 30 septembre-4 octobre 2002, Le Bono, Morbihan** / ouvrage coordonné par Jean-Claude Le Moal, Bernard Hidoine et Lisette Calderan. — Paris : ADBS éd., 2002. — 322 p. — (Collection Sciences de l'information. Série Études et techniques, ISSN 1160-2376). — Notes bibliogr. — ISBN 2-84365-062-3

La problématique de la recherche d'information sur les réseaux est abordée de points de vue différents par des auteurs exerçant des métiers liés à l'information : professionnels de l'information-documentation, informaticiens, consultants, chercheurs, enseignants... Sont abordés plus particulièrement : les instruments de recherche sur le web, les techniques de traitement automatique des langues, la veille stratégique sur les réseaux.

[3] IDC, 2001

💻 **The High Cost of Not Finding Information** [en ligne] / IDC. — [consulté le 13 octobre 2003]
< www.knowledge-wave.com/scripts-include/ en-us/downloads/idcinfo2996.pdf >

Dresse un constat chiffré inquétant des pertes d'argent dues à l'incapacité de retrouver l'information dans les entreprises.

[4] Lefèvre, 2000

📖 **La recherche d'informations : du texte intégral au thésaurus** / Philippe Lefèvre. — Paris : Hermès science publications, 2000. — 253 p. — ISBN 2-7462-0173-9

Fait le point sur la problématique de la recherche d'informations. La description des caractéristiques du langage naturel et des traitements linguistiques est suivie d'un aperçu des langages documentaires, des modes d'analyse et d'indexation, des techniques et modes de requête et d'une caractérisation des moteurs de recherche.

GESTION DOCUMENTAIRE, LANGAGES DOCUMENTAIRES

[5] Maniez, 2002

📖 **Actualité des langages documentaires : fondements théoriques de la recherche d'information** / Jacques Maniez. — Paris : ADBS éd., 2002. — 395 p. — (Collection Sciences de l'information. Série Études et techniques, ISSN 1160-2376). — Bibliogr. p. 365-373. Index. — ISBN 2-84365-060-7

Se propose de préciser le rôle et l'importance des langages documentaires. Contient un bref historique de la gestion documentaire au cours du XXe siècle.

[6] AFNOR, 1987

📖 **Vocabulaire de la documentation** / AFNOR [Association française de normalisation]. — 2e éd. — Paris-La Défense : AFNOR, 1987. — 159p. — (Les Dossiers de la normalisation, ISSN 0297-4827). — ISBN 2-12-484221-8

Édition en vigueur du glossaire qui fournit les définitions officielles des termes de l'information-documentation.

[7] Le Coadic, 1997

📖 **Dictionnaire encyclopédique de l'information et de la documentation** / [réd. par] Yves F. Le Coadic, Michel Melot, Paul-Dominique Pomart [et al.] ; [sous la dir.] de Serge Cacaly. — [Paris] : Nathan, 2001. — 634 p. — ISBN 2-09-191252-2

Définit les concepts, les techniques et les méthodes de l'information-documentation.

[8] Dégez, 2001

📖 **Thésauroglossaire des langages documentaires : un outil de contrôle sémantique** / Danièle Dégez, Dominique Ménillet. — Paris : ADBS éd., 2001. — 181 p. — (Collection Sciences de l'information. Série Recherches et documents, ISSN 1159-7666). — Bibliogr. — ISBN 2-84365-051-8

Thésaurus des langages documentaires accompagné d'un glossaire et d'un lexique français-anglais.

[9] EBSI, 2002

💻 **Terminologie de base en sciences de l'information : volets 1 et 2** [en ligne] / Université de Montréal, Faculté des arts et des sciences, École de bibliothéconomie et des sciences de l'information. — Montréal : Université de Montréal, mise à jour le 16 octobre 2002 [consulté le 27 septembre 2003]
< http://www.ebsi.umontreal.ca/termino/index.htm >

Terminologie de l'information-documentation mise en ligne sur le serveur de l'EBSI.

INDEXATION MANUELLE ET INDEXATION AUTOMATIQUE

[10] Jolion, 2000

📖 **L'Indexation** / sous la dir. de Jean-Michel Jolion. — [Paris] : [Hermes sciences publications], 2000. — 182 p. — No de : "Document numérique", ISSN 1279-5127, vol. 4, no. 1-2, 2000. — ISBN 2-7462-0224-7

Numéro spécial de la revue "Document numérique" centré sur la problématique de l'indexation.

[11] Metzger, 2001

📖 **L'indexation à l'ère d'Internet : actes du congrès d'ISKO-France, École nationale supérieure des sciences de l'information et des bibliothèques et Université Jean Moulin Lyon-3, Lyon, 21-22 octobre 1999** / éd. Jean-Paul Metzger, Mohamed Hassoun, Omar Larouk. — [Paris] : ISKO-France ; Villeurbanne : ENSSIB, 2001. — 240 p. — Notes bibliogr. — ISBN 2-910227-39-1

Rassemble les actes du colloque ISKO-France, tenu en octobre 1999 à Lyon, dont le thème central, l'indexation, se déclinait en cinq grandes thématiques : accès aux ressources sur Internet, espace documentaire, indexation d'images, approches sémantiques et discoursives et méthodes d'indexation.

[12] Amar, 2000

📖 **Les fondements théoriques de l'indexation : une approche linguistique** / Muriel Amar. — Paris : ADBS éd., 2000. — 355 p. — (Sciences de l'information. Recherches et documents, ISSN 1159-7666). — Bibliogr. p. 335-348. — Th. doct. : Sci. information et communication : Lyon 2 : 1997. — ISBN 2-84365-042-9

S'attache à dégager les caractéristiques de l'indexation en étudiant ses fondements théoriques à travers différents modèles linguistiques qui ont contribué à l'évolution des technologies. Aborde premièrement le problème du lexique, ensuite celui de la référence et construit, dans un deuxième temps, un modèle d'utilisation de la langue en indexation.

1.2 LES SYSTÈMES D'ORGANISATION DES CONNAISSANCES

[13] Soergel, 2003

From legacy KOS to full-fledged ontologies [en ligne] / Dagobert Soergel ; Kathy Newton. — Kent, Ohio : Network Knowledge Organization Systems, last updated 16 may 2003 [consulté le 27 septembre 2003]
< http://nkos.slis.kent.edu/2003workshop/soergel.ppt >

Conseils concernant la conversion de langages documentaires classiques en ontologies.

[14] Smith, 2003

Concept-based Learning Spaces : Apply domain-specific KOS principles for organizing collections/services for given applications [en ligne] / Terence M. Smith ; Marcia Zeng. — Kent, Ohio : Network Knowledge Organization Systems, last updated 16 may 2003 [consulté le 27 septembre 2003]
< http://nkos.slis.kent.edu/2003workshop/zengnkos.ppt >

Présentation des langages contrôlés dans le contexte de la recherche dans des bases de connaissances, lors de la 6e conférence NKOS (Network Knowledge Organization Systems/Services).

[15] Adams, 2002

The Semantic Web : Differentiating Between Taxonomies and Ontologies / Katherine Adams. — Online, July-August 2002, vol. 26, no. 4, p. 20-23

Le caractère « sémantique » du Web est de plus en plus accentué. L'article traite des outils mis en oeuvre et indique les tendances les plus productives. Parmi les "outils" hiérarchisés, les taxonomies dans leur emploi par les professionnels de l'information (plutôt que les systématiciens) et les ontologies dans celui des scientifiques (plutôt que des philosophes) sont pointés et caractérisés. La préoccupation de facilitation d'accès à l'information et l'approche du mode de compréhension de l'usager des professionnels de l'information qui produisent des taxonomies constituent une différence cruciale avec les constructeurs d'ontologies. Traduits en terme d'outillage de recherche d'information sur le web, les ontologies mettent en jeu des automates à règles d'inférence existant en intelligence artificielle susceptibles de recourir aux métadonnées, éléments inutiles dans les taxonomies. Le futur est pointé par le développement des "topic maps", comme des taxonomies super-sophistiquées, que définit la norme ISO 13250 et que l'OCLC envisage d'utiliser sur le Web. Ce type d'outil, indépendant des documents qui lui sont liés, serait réutilisable et partageable par des groupes d'utilisateurs et des organisations divers.

[16] Ramos, 2002

Taxonomy, Thesaurus, Tagging : Balancing Automation and Editorial Review / Laura Ramos. — Planning Assumption, Giga Information Group, 8 january 2003

L'analyste du Giga Group présente les taxonomies, thésaurus, ontologies et leurs enjeux dans les entreprises.

TAXONOMIES

[17] Farnum, 2002

Redesigning an e-business taxonomy : Egreetings project case study / Chris Farnum. — Bulletin of the American Society for Information Science and Technology, June-July 2002, vol. 28 no. 5, p. 10-13

L'auteur a participé à la rénovation du site "Egreetings" d'envoi de cartes de voeux, de salutation, avec pour but entre autres, d'augmenter le nombre d'envois. Egreetings était généralement placé au troisième rang des sites web de salutation. Il s'agissait aussi d'améliorer la navigation et la recherche, trouver des actions de promotion, trouver une idée de collection de salutations musicales, améliorer le procédé de contrôle. Ces cartes sont des images, animées ou non, et il fallait faire un effort de conception pour renouveler la collection; des artistes graphiques de l'équipe s'y sont employés.

[18] Bertolucci, 2003

Happiness Is Taxonomy : Four Structures for Snoopy / Katherine Bertolucci. — Information Outlook, March 2003, vol. 7 no. 3, p. 36-44

Katherine Bertolucci est "manager d'information", et aime créer de nouveaux accès à la connaissance, et concevoir des taxonomies (organisation de l'information en catégories hiérarchiques). Elle présente ici l'art et la manière de créer une taxonomie adaptée au sujet et à l'utilisateur à l'aide de l'exemple d'un personnage de bandes dessinée : Snoopy. Cet exemple lui permet une comparaison rapide entre les classifications existantes (la classification de Linné, celles de Dewey et de la Library of Congress)et la taxonomie spécifique qu'elle a créé pour la société Determined Productions qui commercialise les produits dérivés des personnages de Charles Schulz.

[19] Ainsbury, 2002

Cataloging's Comeback : Classifying and Organizing Corporate Documents / Bob Ainsbury. — Online, March-April 2002, vol. 26, no. 2, p. 27-31

Les recherches d'informations qui n'aboutissent pas sont contreproductives pour les entreprises. Fournir au personnel de celles-ci des outils de recherche efficaces passent par un accord dans l'entreprise sur la façon de nommer ce qui est recherché. Les moyens universels sont les moins productifs car ils ne prennent en compte ni la spécificité de l'entreprise, ni les points de vue différenciés à l'intérieur de celle-ci. La mise en place de taxonomies d'entreprises est un investissement permettant d'améliorer la vitesse de recherche et la pertinence de l'information retrouvée. Les métadonnées standards qui font partie de la description des documents ne sont pas destinées à cet usage. La construction d'une taxonomie doit être pensée spécialement à cet effet. Pour les très grandes quantités d'informations à qualifier, la solution réside dans la recherche automatique par des algorithmes de séquences de mots, de co-occurences, de similitudes et de différences de documents pour extraire des concepts et créer des groupements (cluster) recevant une catégorisation. Un procédé interactif de création de cluster consiste à construire un modèle de catégorisation et à le faire fonctionner automatiquement. La réponse donne des indications sur sa validité. Une autre manière, consiste à construire des règles de recherches déployées informatiquement pour construire la taxonomie. Cette solution est coûteuse en efforts humains mais les règles servent ultérieurement en requêtes. Ces procédés ne sont pas exclusifs les uns des autres et peuvent être panachés. L'auteur rappelle qu'en définitive, une taxonomie est faite pour ses utilisateurs et que de la même manière qu'une entreprise doit connaître ses clients, elle doit connaître les utilisateurs de sa ou ses taxonomies.

[20] Saeed, 2002

Using Dewey decimal classification scheme (DDC) for building taxonomies for knowledge organisation / Hamid Saeed ; Abdus Sattar Chaudhry. — Journal of Documentation, 2002, vol. 58, no. 5, p. 575-583

Projet d'application de la classification décimale Dewey afin de faciliter la navigation à travers le web. L'article décrit la 1ère phase de construction de la taxonomie servant à définir des catégories, à partir de la classification Dewey et du thésaurus IEEE du web. La 2ème phase portera sur l'utilisation et l'évaluation de cette taxonomie.

[21] Ramos 2003

Best practices in Taxonomy Development and Management / Laura Ramos ; Daniel W. Rasmus. — Planning Assumption, Giga Information Group, 8 January 2003

Les deux analystes du Giga présentent les enjeux des taxonomies dans les entreprises.

[22] Gilchrist, 2001

Corporate taxonomies : report on a survey of current practice / Alan Gilchrist. — Online Information Review, 2001, vol 25, no. 2, p. 94-102

Cette étude s'intéresse aux pratiques actuelles de construction et d'utilisation de taxonomies, principalement dans les grosses entreprises, mais aussi dans les applications des portails accessibles au public. Comme il n'existe pas de définition de départ de la taxonomie des mots, cette étude a découvert un large éventail de techniques déployées pour aborder le problème, dont toutes sont apparues valables. Trois avantages découlent de cette approche : d'abord elle aide à éclaircir les relations et les différences qui existent entre classifications et thésaurus d'une part et taxonomie de l'autre ; ensuite, elle suggère la possibilité de l'existence d'une évolution naturelle vers l'utilisation d'une approche taxonomique ; enfin elle montre qu'une définition de la taxonomie est en train de naître, dans laquelle la structure (classification) et l'étiquetage (thésaurus) entrent en jeu. Six des vingt-deux études de cas sur lesquels repose cette étude sont ici brièvement présentées afin d'illustrer les différentes approches adoptées et leurs divers aspects.

[23] Knox, 2003

Taxonomy Development : Build or Buy ? / R. Knox, K. Harris, F. Caldwell, D. Logan. — Research Note, Giga Information Group, 9 September 2003

Les analystes du Giga Group posent la question de l'achat de taxonomies, en concluant que cela reste à juger au cas par cas.

THÉSAURUS

[24] Chaumier, 2003b

📖 **Les techniques documentaires au fil de l'histoire, 1950-2000** / Jacques Chaumier ; en collab. avec Florence Gicquel. — Paris : ADBS éd., 2003. — 179 p. — (Collection Sciences de l'information. Série Études et techniques, ISSN 1160-2376). — Bibliogr. p. 133-139. Index. — ISBN 2-84365-064-X

Présente l'évolution des outils et des méthodes utilisés par la profession au cours de la seconde moitié du XXe siècle.

[25] Dextre Clarke, 2003

💻 **BS 8723 : a new British Standard for structured vocabularies** [en ligne] / Stella G. Dextre Clarke. — [consulté le 30 septembre 2003]
< http://www.glam.ac.uk./soc/research/hypermedia/NKOS-workshop%20Folder/dextre_clarke.ppt >

Intervention présentée à la Conférence NKOS qui a eu lieu le 21 août 2003 à Trondheim, en Norvège. Fait le point sur l'état d'avancement des travaux du groupe chargé de la rédaction de cette nouvelle norme britannique.

[26] Warner, 2003

💻 **Guidelines and Principles for Developing Search and Browse Vocabularies** [en ligne] / Amy J. Warner. — Kent, Ohio : Network Knowledge Organization Systems, last updated 16 may 2003 [consulté le 27 septembre 2003]
< nkos.slis.kent.edu/2003workshop/amy.ppt >

Intervention de la consultante en gestion de langages documentaires (Information Architect) américaine Amy Warner lors de la 6e conférence NKOS, le 31 mai 2003 à Houston, Texas. Présente à la fin les principes directeurs des travaux en cours pour la révision de la norme ANSI/NISO qui régit les thésaurus.

[27] Hudon, 1994

📖 **Le Thésaurus : conception, élaboration, gestion** / par Michèle Hudon. — Québec : Éd. ASTED, 1994. — 220 p. — (Clé en main). — Bibliogr. p. 212-213. Index. — ISBN 2-921548-14-3

Présente les thésaurus de descripteurs, les normes régissant la construction des thésaurus monolingues et multilingues et les étapes de construction. Donne de nombreux exemples issus de l'expérience de l'auteur.

TOPIC MAPS

[28] Mondeca, 2002

💻 **Making sense of content** [en ligne] / Mondeca. — Paris, Mondeca, 2003 [consulté le 5 octobre 2003]
< http://www.mondeca.com/english3/published-doc/ITM-doc8p-fr.pdf >

Présente la solution logicielle Intelligent Topic Manager de la société Mondeca (Livre blanc de l'éditeur).

[29] ISO/IEC, 2002

💻 **ISO-IEC 13250 Topic Maps : Information Technology : Document Description and Processing Languages** [en ligne] / ISO ; IEC. — [consulté le 5 octobre 2003]
< www.y12.doe.gov/sgml/sc34/document/ 0322_files/iso13250-2nd-ed-v2.pdf >

Norme des Topic Maps, version du 19 mai 2002. Décrit les Topic Maps et leur expression en HyTM et XTM.

[30] Rath, 2003

💻 **The Topic Maps Handbook : White Paper** [en ligne] / H. Holger Rath. — Gutersloh, Germany : Empolis, Arvato Knowledge Management, 2003 [consulté le 16 août 2003]
< http://www.empolis.com/download/docs/whitepapers/empolistopicmapswhitepaper_eng.pdf >

L'auteur, qui a participé au commité responsable de la norme Topic maps, présente les Topic Maps, leur structure et la fçon de les créer.

[31] Pepper, 2000

💻 **The TAO of Topic Maps : Finding the Way in the Age of Infoglut** [en ligne] / Steve Pepper. — [consulté le 7 octobre 2003]
< www.gca.org/papers/xmleurope2000/pdf/s11-01.pdf >

Papier présenté par cet ingénieur de la société Ontopia à la conférence XML Europe 2000. Précise la construction des Topic Maps.

ONTOLOGIES

[32] Charlet, 2000

📖 **Ingénierie des connaissances : évolutions récentes et nouveaux défis** / Jean Charlet, Manuel Zacklad, Gilles Kassel... [et al.]. — Paris : Eyrolles, 2000. — X-610 p. — (Collection technique et scientifique des télécommunications). — Notes bibliogr. Index. — ISBN 2-212-09110-9

Présente l'ingéniérie des connaissances, discipline jeune et pluridisciplinaire qui étudie les concepts, méthodes et techniques qui permettent de modélise et/ou d'acquérir des connaissances. Les 35 articles ont été sélectionnés parmi ceux présentés dans différents manifestations entre 1995 et 1998. Ceux regroupés dans le 3e chapitre s'intéressent aux ontologies, notamment une contribution de Bruno Bachimont.

[33] Gruber, 1993

A Translation Approach to Portable Ontology Specifications [en ligne] / Thomas R. Gruber. — Knowledge Acquisition, 1993, no. 5, p. 199-220
< http://www.ksl.Stanford.EDU/knowledge-sharing/papers/ontolingua-intro.rtf >

Décrit l'ontologie Ontolingua et fournit des précisions quant aux relations gérées par les ontologies.

[34] Gruber, 1995

Toward Principles for the Design of Ontologies Used for Knowledge Sharing [en ligne] / Thomas R. Gruber. — International Journal of Human and Computer Studies, 1993, no. 43, p. 907-928
< http://ksl-web.stanford.edu/knowledge-sharing/papers/onto-design.rtf >

Formalise les critères de construction d'une ontologie.

[35] Ding, 2002a

Ontology research and development. Part 1 : a review of ontology generation / Ding Ying ; Schubert Foo. — Journal of information science, 2002, vol. 28 no. 2, p. 123-136

L'ontologie est un champ multidisciplinaire complexe, qui s'intéresse à la connaissance, au management et à la compréhension de l'organisation de l'information. Son rôle est important, qui cherche à fournir qualitativement de nouveaux niveaux de services dans la prochaine transformation du web, sous la forme d'un web sémantique. Cette étude est présentée en deux parties : la première concerne l'état de l'art dans les techniques.

[36] Vickery, 1997

Ontologies / B. C. Vickery. — Journal of information science, 1997, vol. 23, no. 4, p. 277-286

L'ontologie est définie par le dictionnaire comme étant «la partie de la métaphysique qui s'applique à l'être, en tant qu'être, indépendamment de ses déterminations particulières». Ce terme apparaît depuis quelques temps dans la littérature en ingénierie de la connaissance - et aussi en science de l'information : dans ce contexte, il a le sens du caractère explicite d'un «monde» qui doit être représenté dans un système informatique. Le contexte qui a conduit à l'émergence de ce concept est évoqué, puis des exemples d'ontologies sont présentés. Le processus de construction d'une ontologie est examiné ainsi que les utilisations de tels outils en ingénierie de la connaissance. En conclusion, l'auteur compare les ontologies avec un certain nombre d'outils similaires utilisés en science de l'information. (classification, thésaurus, lexicographie).

[37] Noy, 2001

Ontology development 101 : A Guide to Creating Your First Ontology [en ligne] / Natalya F. Noy ; Deborah L. McGuinness. — Stanford, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001 [consulté le 27 juillet 2003]
< http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf >

Présente les ontologies et les méthodes de construction possibles (utilisant les logiciels Protégé-2000, Ontolingua, Chimaera). Les exemples données utilisent Protégé-2000 et montrent comment construire une ontologie des vins qui peut avoir des applications dans la restauration.

[38] Bachimont, 1995

Ontologie régionale et terminologie : quelques remarques critiques et méthodologiques / B. Bachimont. — La Banque des mots, no. spécial 7-1995, Terminologie et intelligence artificielle, p. 67-86

Pour construire un système à base de connaissances (SBC), l'auteur rappelle les deux problèmes essentiels de l'acquisition des connaissances : cerner les objets ou notions mobilisés dans un domaine et en donner une description compatible avec une exploitation informatique. Il propose une méthodologie de construction d'« ontologie régionale », ou théorie conceptuelle d'un domaine.

[39] Gandon, 2002

Distributed artificial intelligence and knowledge management : ontologies and multi-agent systems for a corporate semantic web [en ligne] / Fabien Gandon. — Nice : INRIA, 2002. — Thèse de doctorat en STIC soutenue en novembre 2002 [consulté le 3 août 2003]

L'auteur de cette thèse a participé au projet européen CoMMA (constituer une ontologie). Titre français : "Intelligence artificielle distribuée et gestion des connaissances : ontologies et systèmes multi-agents pour un web sémantique organisationnel".

[40] Zweigenbaum, 1996

Le rôle du lexique sémantique et de l'ontologie dans le traitement automatique de la langue médicale [en ligne] / Pierre Zweigenbaum ; Bruno Bachimont ; Jacques Bouaud,... [et al.]. — [consulté le 25 août 2003]
< http://www-test.biomath.jussieu.fr/~pz/FTPapiers/Zweigenbaum:CRISTALS96.pdf >

Présente les ontologies, spécialement du domaine médical, notamment l'ontologie Menelas.

[41] Dechilly, 2000

Une ontologie pour éditer des schémas de description audiovisuels, extension pour l'inférence sur les descriptions [en ligne] / Thomas Dechilly ; Bruno Bachimont. Paris, INA, 2000 [consulté le 12 septembre 2003]
< http://www.irit.fr/IC2000/ACTES/DechillyIC00.pdf >

Présente l'ontologie du domaine audiovisuel, en cours de développement à l'INA, liée à la norme MPEG-7.

[42] Ding, 2002b

Ontology research and development. Part 2 : a review of ontology mapping and evolving / Ding Ying ; Schubert Foo. — Journal of information science, 2002, vol. 28 no. 2, p. 375-388

Cette 2e partie sur l'ontologie traite de sa cartographie et de son évolution. L'ontologie doit être capable d'intégrer les changements de relations et de significations. Sa cartographie concerne la réutilisation, l'extension et la combinaison des ontologies existantes. Son évolution comprend la maintenance des ontologies existantes et leur adaptation en fonction des nouvelles informations. Selon les auteurs, les travaux en ontologie semi-automatique et automatique couvrant les 3 aspects, création, cartographie et évolution ne sont qu'une percée au succès limité. L'intervention d'un expert humain est essentielle dans la plupart des cas. L'apport de l'ontologie dans le web sémantique est discuté.

1.3 LES MOTEURS DE RECHERCHE ET LEUR FONCTIONNEMENT

[43] Leloup, 1998

📖 **Moteurs d'indexation et de recherche : environnements client-serveur, Internet et intranet** / Catherine Leloup. — Paris : Eyrolles, 1998. — XIII-285 p. + 1 CD-ROM. — Bibliogr. Glossaire. Lexique français-anglais et anglais-français. Index. — ISBN 2-212-08976-7

Quelque peu ancien en ce qui concerne les produits et les services ainsi que les sociétés éditrices ou commerciales, ce livre reste d'actualité en ce qui concerne le mode de fonctionnement des moteurs de recherche. Une première partie est consacrée à l'explication des traitements statistiques et linguistiques ainsi que des concepts qui sont à la base de la recherche par le contenu. Dans un deuxième temps, l'offre logicielle disponible en 1997 est analysée et testée.

1.3.1 MOTEURS DE RECHERCHE ET TRAITEMENT AUTOMATIQUE DES LANGUES « NATURELLES » (TALN)

[44] Fuchs, 1993

📖 **Linguistique et traitements automatiques des langues** / par Catherine Fuchs ; avec la collab. de Anne Lacheret-Dujour et de Bernard Victorri. — [Paris] : Hachette supérieur, 1993. — 303 p. — (HU. Linguistique) (Hachette université. Langue, linguistique, communication). — Bibliogr. Index. — ISBN 2-01-016908-5

Les auteurs, chercheurs au CNRS, présentent le traitement automatique des langues qui s'appuie sur des concepts, des modèles et des analyses linguistiques.

[45] Rastier, 1994

📖 **Sémantique pour l'analyse : de la linguistique à l'informatique** / François Rastier, Marc Cavazza, Anne Abeillé. — Paris ; Milan ; Barcelone : Masson, 1994. — XII-240 p. — (Sciences cognitives, ISSN 0991-577X). — Bibliogr. p. 225-234. Index. - ISBN 2-225-84537-9

Publié en 1994, l'ouvrage collectif s'appuie sur des articles plus anciens et propose un aperçu de la sémantique, de ses méthodes et des ses relations avec l'informatique. Le premier et le deuxième chapitre, écrits par F.Rastier, font le point sur la façon dont la théorie sémantique est utilisée dans les applications informatiques. Un épilogue présente les directions de recherche et les applications industrielles attendues.

[46] Habert, 1997

📖 **Les linguistiques de corpus** / Benoît Habert, Adeline Nazarenko, André Salem. — Paris : A. Colin, 1997. — 240 p. — (Collection U. Série Linguistique). — Bibliogr. Index. — ISBN 2-200-01775-8

Propose une vue d'ensemble des travaux récents dans le domaine du traitement automatique des textes. Les méthodes d'analyse linguistique appliquées à un corpus de textes sont décrites et illustrées d'exemples issus de travaux de recherche.

[47] Chowdhury, 2003

Language and representation : 2) Natural Language Processing / Gobinda G. Chowdhury. — Annual review of information science and technology, 2003, vol. 37, p. 51-90

Le Traitement du langage naturel (NLP)est un secteur de recherche et d'application qui explore comment les ordinateurs peuvent être employés pour comprendre et manipuler le texte ou le discours en langage naturel. Les chercheurs en NLP tentent d'expliquer comment les gens comprennent et emploient le langage, afin de pouvoir développer des outils et des techniques appropriés pour concevoir des systèmes d'information capables de comprendre et de manipuler les langages naturels pour exécuter des tâches. Les bases du traitement du langage naturel se trouvent dans un certain nombre de disciplines : les sciences de l'information, l'informatique, la linguistique, les mathématiques, l'ingénierie électrique et électronique, l'intelligence artificielle, la robotique et la psychologie. Ses applications incluent de nombreux domaines, comme la traduction automatique, le résumé de texte, des interfaces utilisateur, la recherche documentaire bilingue (CLIR : cross-language information retrieval) et multilingue, la reconnaissance de la parole, l'intelligence artificielle et les systèmes experts. Un secteur d'application important et relativement nouveau concerne le World Wide Web et les bibliothèques numériques. Plusieurs chercheurs ont souligné le besoin de recherche appropriée pour faciliter la recherche documentaire bilingue ou multilingue (y compris le traitement de textes multilingues et des systèmes d'interface utilisateur multilingues) pour exploiter au mieux les bibliothèques numériques. Fait le point sur ces domaines d'application et de recherche.

[48] Mahmoudi, 1997

Traitement automatique des langues naturelles : évolution et perspectives / Seyed Mohammad MAHMOUDI. — Revue de la science de l'information, juillet 1997, no. 6, p. 9-60

Le traitement automatique des langues naturelles (TALN) est au carrefour de disciplines variées telles que la linguistique, les mathématiques, les statistiques et les sciences cognitives. Le TALN recourt notamment à la logique comme mécanisme de représentation déductive et emprunte ses méthodes et outils aux applications d'intelligence artificielle en matière de représentation du savoir et d'élaboration des logiciels et analyseurs (morpho-syntaxique et sémantique). Le TALN constitue ainsi un champ de prédilection pour le développement simultané de l'ensemble des disciplines associées. Les domaines d'application du TALN sont par ailleurs multiples ; l'article développe principalement les aspects liés à l'indexation automatique et à la traduction automatique. Le traitement automatique des langues naturelles est en perpétuel développement ; les systèmes de TALN restent cependant perfectibles. L'article fait état des problèmes qui demeurent encore non résolus.

[49] SATO, 2001

Glossaire des termes d'ATO [en ligne] / Université du Québec à Montréal, Service ATO ; École de bibliothéconomie et des sciences de l'information ; Documensa. — Montréal : Université du Québec à Montréal, [consulté le 27 septembre 2003]
< http://www.ling.uqam.ca/sato/glossaire/index.html >

Glossaire concernant l'analyse de textes par ordinateur (ATO).

[50] OTIL, 2003

Lexique OTIL [en ligne] / Observatoire du traitement informatique des langues et de l'inforoute. — Liège : OTIL, dernière modification 1 août 2003 [consulté le 27 septembre 2003]
< http://www.owil.org/fr_lexique.htm >

Glossaire de termes concernant le traitement informatique de la langue (TIL).

[51] Sabah, 1989

📖 **L'intelligence artificielle et le langage. 2, Processus de compréhension** / Gérard Sabah. — Paris : Hermès, 1989. — 411 p. — (Langue, raisonnement, calcul, ISSN 0988-0569). — Bibliogr. p. 375-395. Index. — ISBN 2-86601-187-2

Le premier volume ayant présenté les modèles linguistiques, les travaux des informaticiens et des psychologues afin de modéliser le langage humain, ce deuxième volume developpe les diférents mécanismes informatiques utilisés en TALN.

[52] Marchand, 1998

📖 **L'analyse du discours assistée par ordinateur : concepts, méthodes, outils** / Pascal Marchand. — Paris : A. Colin, 1998. — 222 p. — (U. Psychologie). — Bibliogr. Notes bibliogr. Index. — ISBN 2-200-21914-8

Une première partie décrit les outils conceptuels d'analyse automatique du discours (langue et contexte) : analyses para-verbales, lexicales, morpho-syntaxiques, sématiques et pragmatiques. La deuxième partie, applicative, propose l'analyse d'un même corpus politico-médiatique au moyen de plusieurs logiciels.

[53] Pierrel, 2000

📖 **Ingénierie des langues** / sous la dir. de Jean-Marie Pierrel. — Paris : Hermès science publications, 2000. — 354 p. — (Traité IC2 : information, commande, communication. Série Informatique et systèmes d'information). — Notes bibliogr. Index. — ISBN 2-7462-0113-5

Recueil d'études qui aborde le traitement automatique des langues (TAL) dans la perspective de la mise en place d'applications concrètes : indexation et accès à l'information, résumé de textes, traduction assistée par ordinateur, correcteurs d'orthographe, dialogue homme-machine etc.

[54] Silberztein, 1993

📖 **Ingénierie des langues** / sous la dir. de Jean-Marie Pierrel. — Paris : Hermès science publications, 2000. — 354 p. — (Traité IC2 : information, commande, communication. Série Informatique et systèmes d'information). — Notes bibliogr. Index. — ISBN 2-7462-0113-5

Recueil d'études qui aborde le traitement automatique des langues (TAL) dans la perspective de la mise en place d'applications concrètes : indexation et accès à l'information, résumé de textes, traduction assistée par ordinateur, correcteurs d'orthographe, dialogue homme-machine etc.

[55] Memodata, 1999

💻 **Memodata** [en ligne] / Memodata. — Caen : Memodata, 10 sept. 1999, dernier ajout octobre 2003 [consulté le 22 octobre 2003]
< http://www.memodata.com/index.shtml >

Site de la société Memodata de Caen. Présente les dictionnaires électroniques commercialisés pour être intégrés à des applications (OEM).

[56] Kayser, 2001

📖 **Traitement automatique du langage naturel** / sous la dir. de Daniel Kayser et Bernard Levrat. — Paris : Hermès, 2001. — p. 249-494. — Notes bibliogr. — No. de "TSI. Technique et science informatiques", ISSN 0752-4072, vol. 20, no. 3, 2001. — ISBN 2-7462-0256-5

Présente des études qui apportent des éléments significatifs pour aborder le traitement automatique du langage naturel (TALN) sous ses différents aspects : formalisme syntaxique (modélisation des phénomènes linguistiques), structure de plate-forme logicielle (modularité), résumé automatique (extraction d'information basée sur des repères linguistiques), classification thématique (appliquée à des résultats de recherche effectuées via Internet), traitement de la syntaxe (techniques de satisfaction de contraintes).

[57] Minel, 2002

📖 **Filtrage sémantique : du résumé automatique à la fouille de textes** / Jean-Luc Minel. — [Paris] : Hermès science publications ; [Cachan] : Lavoisier, 2002. — 202 p. — Bibliogr. p. 187-199. Index. — ISBN 2-7462-0602-1

Présente les méthodes qui permettent de fournir l'information pertinente extraite des masses importantes de documents textuels et insiste sur la nécessité d'utiliser la sémantique pour des meilleurs drésultats dans la fouille de textes.

[58] Chaudiron, 2001

📖 **Filtrage et résumé automatique de l'information sur les réseaux : actes du 3e Congrès du chapitre français de l'ISKO, Nanterre, 5-6 juillet 2001** / éd. Stéphane Chaudiron et Christian Fluhr. — Nanterre, Université de Nanterre Paris X, 2001. — 283 p. — Index. — ISBN 2-9516737-0-1

Actes du 3e Congrès du Chapitre français de l'ISKO tenu à Paris en juillet 2001, consacré au filtrage sémantique et aux techniques de résumé automatique.

[59] Jacquemin, 2000

📖 **Traitement automatique des langues pour la recherche d'information** / sous la dir. de Christian Jacquemin. — Paris : Hermès, 2000. — p. 327-591. — Notes bibliogr. — No de : "TAL. Traitement automatique des langues", ISSN 1248-9433, vol. 41, no. 2, 2000. — ISBN 2-7462-0225-5

Études sur le traitement automatique des langues (TAL) pour la recherche d'information (RI). Tend à montrer que le TAL a atteint actuellement une maturité technologique suffisante pour traiter finement l'information contenue dans des masses importantes de données textuelles. S'articule autour de trois thèmes majeurs : premièrement, la constitution d'outils et de ressources pour le TAL appliqué à la RI, deuxièmement, la description d'applications pour l'accès à l'information reposant sur des méthodes de TAL, enfin, l'enrichissement d'outils de RI par des techniques de TAL. Sont présentés : l'acquisition de liens lexicaux transcatégoriels, un analyseur statistique pour l'extraction automatique des unités lexicales complexes, un système reposant sur des analyses syntaxiques et sémantiques et sur un algorithme de démonstration pour la recherche de réponses. Sont proposées aussi : deux possibilités d'enrichir l'indexation des documents (par une dimension sémantique qui repose sur des fréquences de cooccurrences et une autre au moyen d'analyseurs morphologiques et syntaxiques à large couverture), enfin, une technique de classification des documents trouvés par un moteur de recherche.

1.3.2 POSITIONNEMENT DES ACTEURS SUR LE MARCHÉ ET TYPOLOGIE DES PRODUITS

MARCHÉ DES MOTEURS DE RECHERCHE

[60] Roberget, 2003

Le marché impose son modèle économique à la recherche sur internet [en ligne] / Olivier Roberget. — 01 Informatique, 5 septembre 2003, no. 1735, p. 6-7 [consulté le 20 septembre 2003]
< http://www.01net.com/Pdf/01I200309051735006.pdf >

Article qui présente, dans un encart, l'état du marché des moteurs de recherche, leurs résultats financiers ainsi que le positionnement des acteurs.

TYPOLOGIE DES MOTEURS DE RECHERCHE

[61] Chaumier, 2003a

Recherche et analyse de l'information textuelle : tendances des outils linguistiques / Jacques Chaumier ; Martine Déjean. — Documentaliste - Sciences de l'information, février 2003, vol. 40, no. 1, p. 14-24

Fondée sur une enquête menée en 2002 auprès des producteurs d'infologiciels disponibles sur le marché français, cette étude propose une analyse de l'offre actuelle d'outils de recherche et d'analyse d'information textuelle. Les auteurs exposent d'abord en détail les principes de l'approche linguistique qui préside à la conception de la plupart des logiciels d'indexation assistée par ordinateur, puis esquissent une typologie des outils linguistiques existants ainsi que des applications documentaires faisant appel à ces outils. Ils étudient ensuite le contexte économique de production de dix-huit infologiciels dont les principales fonctions sont présentées dans un tableau. Pour finir, ils envisagent les perspectives de développement des outils linguistiques de traitement de l'information

[62] Crochet Damais, 2003

Panorama des outils de recherche [en ligne] / Antoine Crochet Damais. — Journal du Net, 2 octobre 2003 [consulté le 5 octobre 2003]
< http://solutions.journaldunet.com/0310/031002_pano_moteur >

Les moteurs de recherche les plus répandus sur le marché français sont présentés de façon succincte, en fonction de leurs caractéristiques.

[63] Chilotti, 2003

Les moteurs sémantiques toujours plus proches du sens des mots [en ligne] / Sandrine Chilotti. — Le monde informatique, 25 avril 2003, no. 979 [consulté le 5 octobre 2003]
< http://www.weblmi.com/articles_store/979_19/Article_view >

Décrit succintement quelques moteurs de recherche présents sur le marché français qui intègrent des traitements sémantiques.

[64] Dalbin, 2002

Indexation automatique et langage naturel [en ligne] / Sylvie Dalbin. — Rouen, ADBS, 2002 [consulté le 10 octobre 2003]
< http://www.galilo.net/normandie-adbs/actualite/ill/dalbin_rouen2002v2.pdf >

Intervention de la consultante Sylvie Dalbin lors de la journée d'étude ADBS du 5 décembre 2002, à Rouen.

[65] Girardeau, 2002

Analyser un logiciel de GED : les fonctionnalités passées au crible / Nicolas Girardeau. — In : GED et solutions logicielles : guide pratique. — Paris : Archimag, 2002. — 96 p.

Dossier paru dans ce guide pratique spécial GED réalisé par la revue Archimag en octobre 2002. Parmi les fiches techniques, certaines concernent des moteurs de recherche.

[66] Crochet Damais, 2002

Comparez les moteurs de recherche [en ligne] / Antoine Crochet Damais. — Journal du Net, 1 janvier 2002 [consulté le 30 septembre 2003]
< http://solutions.journaldunet.com/0203/020312_moteur.shtml >

Pour permettre de comparer les performances des moteurs de recherche, cet article donne quelques adresses de sites de grandes entreprises où ces technologies ont été mises en oeuvre.

[67] TripleHop Technologies, 2002

Match Point, Information Retrieval Software : Differentiating Factors [en ligne] / TripleHop Technologies. — [consulté le 16 septembre 2003]
< www.triplehop.com/pdf/ MatchPoint%20Competitive%20Advantage%20White%20Paper.pdf >

Livre blanc de l'éditeur du moteur de recherche Match Point.

[68] Villacampa, 1999

La création d'une base de donnés avec Retrievalware d'Excalibur / Alberto Villacampa. — Paris, INTD, 1999. — 43 p. — Mémoire INTD

Le présent mémoire a pour objet de définir des éléments de réponse à cette question : existe-t-il une opposition entre indexation manuelle et indexation automatique ? Il s'appuie sur une mission de stage à Air France qui a consité en la création d'une base de données à l'aide du logiciel Retrievalware de la société Excalibur. Une méthode alternative est proposée qui combine indexation manuelle et automatique.

[69] Dalbin, 2000

Une expérience d'utilisation d'un système d'information documentaire en langage naturel / Sylvie Dalbin ; Bruno Saleras. — Documentaliste - Sciences de l'information, décembre 2000, vol. 37, no. 5-6, p. 312-324

Avec le développement des intranets, de nombreux centres de documentation sont conduits à modifier leur mode d'organisation et de fonctionnement : du modèle "classique" avec un logiciel de gestion et de recherche de références bibliographiques, ils passent à un système d'information en langage naturel directement accessible aux usagers. Retour d'une expérience menée au Centre documentaire des Assurances générales de France (AGF), cette étude voudrait donner au lecteur les moyens d'évaluer les apports et les limites de la recherche en langage naturel, pour les usagers comme pour les documentalistes. Après avoir présenté l'intranet documentaire des AGF et précisé les aspects terminologiques et techniques des systèmes d'indexation et de recherche en langage naturel, cet article étudie les profils et le comportement des usagers, ainsi que l'évolution de leurs pratiques ; et analyse les questions posées et les réponses obtenues. Il met également l'accent sur le nécessaire "décentrage" -par rapport à nos pratiques et habitudes- que suppose, voire qu'impose le développement de tels dispositifs orientés "usager final".

[70] Influo Software, 2002

Moteurs de recherche : où est la technologie ? [en ligne] / Influo Software. — [consulté le 10 septembre 2003]
< http://h2ptm.hymedia.univ-paris8.fr/webmining/prez/Influo%20WP%20moteurs%20V1.pdf >

Livre blanc de l'éditeur Influo Software, technologie issue du CNRS.

[71] Villoing, 2003

Influo Software : serveurs de connaissances, la neuromimétique au service d'un Web sémantique [en ligne] / F. Villoing. — 14 ianvier 2003 [consulté le 15 septembre 2003]
< http://h2ptm.hymedia.univ-paris8.fr/webmining/prez/influo_fichiers/frame.htm >

Présentation du moteur Influo à la journée d'étude "Recherche et fouille d'information sur le Web : perspectives sémantiques et statistiques", tenue à l'Université Paris 8 le 14 janvier 2003.

1.3.3 FONCTIONNEMENT DES MOTEURS DE RECHERCHE LINGUISTIQUES : L'EXEMPLE DE VERITY K2

[72] Roumieux, 2000b

Linguistique : domestiquer l'ordinateur / Olivier Roumieux ; Michel Remize ; Charlotte Perrin. — Archimag, novembre 2000, no. 139, p. 21-31

Les besoins liés à la recherche sur l'Internet, la traduction automatique, le multilinguisme ou l'aide à la décision avec la production de résumés ou de synthèses automatiques ont donné un puissant élan aux outils en langage naturel. La vocation de ces outils est d'apporter de l'intelligence par le biais de traitements linguistiques de différents niveaux : morphologique, syntaxique, sémantique et enfin pragmatique. Mais cette intelligence n'exclut par un certain niveau de compréhension du système par l'utilisateur. Le récent engouement pour ces logiciels et la nouvelle dénomination commerciale de certains d'entre eux pourraient laisser penser, à tort, que leur origine est récente. En réalité, certains font partie du paysage documentaire depuis longtemps comme Lexiquest, Spirit ou Intuition, même si de nombreuses jeunes sociétés se développent comme Albert. Le pragmatisme est à l'origine d'une réorientation technologique actuelle chez les éditeurs : des niches se développent sur des marchés identifiés comme l'indexation à distance ou des versions gratuites pour les sites personnels. Quant à la recherche scientifique, elle se poursuit par exemple au laboratoire TALANA de Paris 7, avec une problématique ancienne : la modélisation des mots de la langue française. L'exemple de la météorologie nationale qui utilise le logiciel Lexiquest pour générer automatiquement des bulletins en fonction du profil de l'utilisateur (un marin ou un citadin) illustre bien l'apport de ces technologies dans les systèmes d'information.

[73] Verity, 2001c

Verity K2 Enterprise, Fundamentals Guide V4.0 / Verity. — Sunnyvale, California : Verity, 2001

Manuel d'utilisation fourni avec le le moteur de recherche Verity K2. Destiné surtout à l'administration (informatique) du moteur. Mais la connaissance de certaines parties est nécessaire à la gestion de VQL. Consulter notamment les parties qui présentent : le fichier « style » qui répertorie les mots vides, les options à prendre pour l'usage de certains opérateurs (ex. : « In ») et la gestion des caractères spéciaux, comme « # » ou « & »). Présente aussi les principes de la taxonomie, de façon synthétique, sous forme de diagrammes.

[74] Verity, 2001e

Verity K2 Search Objects Guide V4.0 / Verity. — Sunnyvale, California : Verity, 2001

Manuel d'utilisation fourni avec le moteur de recherche Verity K2. Contient un schéma de l'architecture de l'application et une présentation générale du langage d'interrogation de Verity (Verity Query Language) ainsi que des parseurs.

[75] Verity, 2001b

Verity K2 Enterprise, Fundamentals Guide V4.0 / Verity. — Sunnyvale, California : Verity, 2001

Manuel d'utilisation fourni avec le le moteur de recherche Verity K2. Destiné surtout à l'administration (informatique) du moteur. Mais la connaissance de certaines parties est nécessaire à la gestion de VQL. Consulter notamment les parties qui présentent : le fichier « style » qui répertorie les mots vides, les options à prendre pour l'usage de certains opérateurs (ex. : « In ») et la gestion des caractères spéciaux, comme « # » ou « & »). Présente aussi les principes de la taxonomie, de façon synthétique, sous forme de diagrammes.

[76] Ambroziak, 2000

Managing tokenizers in XML search [en ligne] / Jacek Ambroziak. — [consulté le 10 octobre 2003]
< http://www.gca.org/papers/xmleurope2000/pdf/s12-04.pdf >

Texte de l'intervention de cet ingénieur de Sun Microsystems à la conférence XML Europe 2000, concernant le découpage de la chaîne de caractères en unités.

[77] Verity, 2001a

Verity Collection Reference Guide V4.0 / Verity. — Sunnyvale, California : Verity, 2001

Manuel d'utilisation fourni avec le moteur de recherche Verity K2. Destiné à l'administration informatique du moteur. Présente les options d'installation et les paramétrges possibles.

[78] Savoy, 1993

Stemming of french words based on grammatical categories / Jacques Savoy. — Journal of ASIS, January 1993, vol. 44, no. 1, p.1-9

Dans les systèmes d'indexation automatique on a souvent recours à des algorithmes qui écartent les suffixes et les préfixes afin de rassembler les mots ayant la même racine sur une même branche. Cet article propose un algorithme qui supprime les suffixes afin d'améliorer ce processus et l'applique à la langue française. Une analyse morphologique est nécessaire pour écarter les suffixes flexionnels ou les variantes morphosyntaxiques d'un lemme (ou proposition préliminaire d'une racine). Après cette analyse un algorithme de suppression des suffixes est implémenté qui utilise un dictionnaire et des catégories grammaticales pour écarter les suffixes. Cette approche permet toujours d'aboutir à un lemme linguistiquement correct, mais qui n'est pas toujours le bon. Le taux d'erreur est d'environ 16 pour cent. L'analyse montre qu'il n'est pas possible d'améliorer les résultats obtenus.

[79] Leloup, 2002

Catégorisation et classification automatiques [en ligne] / Catherine Leloup. — 19 avril 2002 [consulté le 15 août 2003]
< http://www.adbs.fr/uploads/journées/leloup/index.htm >

Intervention de la consultante Catherine Leloup lors de la journée d'étude ADBS du 12 avril 2002. Fait la distinction entre la catégorisation automatique et la classification automatique

[80] Verity, 2001d

Verity K2 Enterprise, Intelligent Classification Guide V4.0 / Verity. — Sunnyvale, California : Verity, 2001

Manuel d'utilisation de l'outil Intelligent Classifier (IC), fourni avec le moteur de recherche Verity K2. Présente l'interface et la marche à suivre pour créer les concepts et la taxonomie à l'aide de l'outil IC ou en ligne de commande. Le volume est fourni en format papier et sous deux formats électroniques : PDF (permettant une consultation aisée à l'écran et la recherche rapide d'une notion qui poserait problème) et HTML (qui permet une consultation plus ludique). Les opérateurs de Verity sont présentés, ainsi que la façon de les combiner.

[81] Verity, 2002

💻 **The ABCs of Content Organization : Verity White Paper** [en ligne] / Verity. — Sunnyvale, California : Verity, 2002 [consulté le 16 septembre 2003]
< www.verity.com/pdf/white_papers/ MK0391a_ContentOrg_WP.pdf >

Livre blanc de l'éditeur Verity. Présente les modalités de gestion des concepts et de la taxonomie, notamment pour la génération automatique.

[82] Verity, 2003

💻 **The Verity K2 Discovery Tier : the Importance of Advanced, Effective Search Tools : Verity White Paper** [en ligne] / Verity. — Sunnyvale, California : Verity, 2003 [consulté le 16 septembre 2003]
< www.verity.co.uk/pdf/white_papers/ MK0348a_Discovery_WP.pdf >

Livre blanc de l'éditeur Verity. Présentation générale des possibilités de l'outil (2003) : les types de recherche possibles (parmi lesquels la recherche suivant la taxonomie), un aperçu rapide des technologies d'indexation et de filtrage, le language de requête propriétaire et, en annexe, les opérateurs.

[83] Verity, 2001f

📖 **Verity Query Language Guide V4.0** / Verity. — Sunnyvale, California : Verity, 2001

Manuel d'utilisation fourni avec le moteur de recherche Verity K2. D'une apparence complexe, le langage d'interrogation de Verity (Verity Query Language ou VQL) est un outil puissant qui permet de construire des requêtes complexes en mode expert. Utilisable autant en interrogation que pour la construction de concepts (Topics) basés sur des requêtes ou de la taxinomie (Taxonomy), le VQL utilise plus de 35 opérateurs. Le volume présente les caractéristiques de chaque opérateur et les règles de combinaison avec d'autres opérateurs. Sa lecture est nécessaire pour mieux comprendre le guide d'utilisation de l'outil Verity Intelligent Classifier.

2 AUDIT : VERITY K2 (CONCEPTS ET TAXONOMIE) ET LE PORTAIL CYBERTHÈQUE

CONTEXTE

[84] Laviale, 1999

✍ **La Cyberthèque de la Direction informatique de la Société Générale : un outil de gestion et de partage des connaissances** / Laure Laviale. — Paris, INTD, 1999. — 91 p. — Mémoire INTD

L'objet du mémoire est de cerner le concept d'outil de gestion et de partage des connaissances et de déterminer si, grâce notamment à un audit, la solution mise en place à la Société Générale (la cyberthèque) fait partie de cette catégorie.

[85] Vacarie, 2001

✍ **Étude comparative des moteurs de recherche sémantique : choix d'un outil pour le service de veille du département informatique de la Société Générale** / Cécile Vacarie. — Paris, INTD, 2001. — 135 p. — Mémoire INTD

Le mémoire fait, dans un premier temps, un point sur les limites des moteurs de recherche en texte intégral et le fonctionnement des moteurs de recherche sémantique. Il s'agit ensuite, à travers un cas concret de choix d'un outil pour la Société Générale, d'élaborer une méthode rigoureuse de comparaison de logiciel ; avant de faire un panorama de l'offre logicielle des plus grands éditeurs.

DÉMARCHE PROJET

[86] Corbel, 2003

🕮 **Management de projet : fondamentaux, méthodes, outils** / Jean-Claude Corbel ; préf. de Pierre-Alain de Smedt. — Paris, Ed. d'Organisation, 2003. — 169 p. — (Les références). — Bibliogr. Index. — ISBN 2-7081-2872-8

Explique les fondamentaux du managemeent de projet, les différentes étapes de déroulement d'un projet, les outils et les méthodes à choisir selon le problème posé, ainsi que les facteurs humains à prendre en compte.

[87] Picq, 1999

🕮 **Manager une équipe projet** / Thierry Picq. — Paris : Dunod, 1999. - VI-225 p. — (Fonctions de l'entreprise. Série Animation des hommes). — La couv. porte en plus : "pilotage, enjeux, performance". - Bibliogr. — ISBN 2-10-004031-6

Présente le fonctionnement en mode projet. Les six premiers chapitres clarifient la notion de projet et d'équipe, les trois suivants montre l'importance de ce type de management dans une démarche de conduite du changement.

[88] Morgat, 1995

📖 **Audit et gestion stratégique de l'information** / Pierre Morgat. — Paris : les Ed. d'Organisation, 1995. — 143 p. — (Collection Audit). — Bibliogr. p. 142-143. — ISBN 2-7081-1826-9

Quelque peu ancien, l'ouvrage propose, au cinquième chapitre, une méthode d'audit de l'information.

BIBLIOGRAPHIE ALPHABÉTIQUE (NOMS D'AUTEURS)

ADAMS, 2003

The Semantic Web : Differentiating Between Taxonomies and Ontologies / Katherine Adams. — Online, July-August 2002, vol. 26, no. 4, p. 20-23

AFNOR, 1987

Vocabulaire de la documentation / AFNOR [Association française de normalisation]. — 2e éd. — Paris-La Défense : AFNOR, 1987. — 159 p. — (Les Dossiers de la normalisation, ISSN 0297-4827). — ISBN 2-12-484221-8

AINSBURY, 2002

Cataloging's Comeback : Classifying and Organizing Corporate Documents / Bob Ainsbury. — Online, March-April 2002, vol. 26, no. 2, p. 27-31

AMAR, 2000

Les fondements théoriques de l'indexation : une approche linguistique / Muriel Amar. — Paris : ADBS éd., 2000. — 355 p. — (Sciences de l'information. Recherches et documents, ISSN 1159-7666). — Bibliogr. p. 335-348. — Th. doct. : Sci. information et communication : Lyon 2 : 1997. — ISBN 2-84365-042-9

AMBROZIAK, 2000

Managing tokenizers in XML search [en ligne] / Jacek Ambroziak. — [consulté le 10 octobre 2003]
< http://www.gca.org/papers/xmleurope2000/pdf/s12-04.pdf >

BACHIMONT, 1995

Ontologie régionale et terminologie : quelques remarques critiques et méthodologiques / B. Bachimont. — La Banque des mots, no. spécial 7-1995, Terminologie et intelligence artificielle, p. 67-86

BERTOLUCCI, 2003

Happiness Is Taxonomy : Four Structures for Snoopy / Katherine Bertolucci. — Information Outlook, March 2003, vol. 7 no. 3, p. 36-44

CHARLET, 2002

📖 **Ingénierie des connaissances : évolutions récentes et nouveaux défis** / Jean Charlet, Manuel Zacklad, Gilles Kassel... [et al.]. — Paris : Eyrolles, 2000. — X-610 p. — (Collection technique et scientifique des télécommunications). — Notes bibliogr. Index. — ISBN 2-212-09110-9

CHAUDIRON, 2001

📖 **Filtrage et résumé automatique de l'information sur les réseaux : actes du 3e Congrès du chapitre français de l'ISKO, Nanterre, 5-6 juillet 2001** / éd. Stéphane Chaudiron et Christian Fluhr. — Nanterre, Université de Nanterre Paris X, 2001. — 283 p. — Index. — ISBN 2-9516737-0-1

CHAUMIER, 2003a

📄 **Recherche et analyse de l'information textuelle : tendances des outils linguistiques** / Jacques Chaumier ; Martine Déjean. — Documentaliste - Sciences de l'information, février 2003, vol. 40, no. 1, p. 14-24

CHAUMIER, 2003b

📖 **Les techniques documentaires au fil de l'histoire, 1950-2000** / Jacques Chaumier ; en collab. avec Florence Gicquel. — Paris : ADBS éd., 2003. — 179 p. — (Collection Sciences de l'information. Série Études et techniques, ISSN 1160-2376). — Bibliogr. p. 133-139. Index. — ISBN 2-84365-064-X

CHILOTTI, 2003

💻 **Les moteurs sémantiques toujours plus proches du sens des mots** [en ligne] / Sandrine Chilotti. — Le monde informatique, 25 avril 2003, no. 979 [consulté le 5 octobre 2003]
< http://www.weblmi.com/articles_store/979_19/Article_view >

CHOWDHURY, 2003

📄 **Language and representation : 2) Natural Language Processing** / Gobinda G. Chowdhury. — Annual review of information science and technology, 2003, vol. 37, p. 51-90

CORBEL, 2003

📖 **Management de projet : fondamentaux, méthodes, outils** / Jean-Claude Corbel ; préf. de Pierre-Alain de Smedt. — Paris, Ed. d'Organisation, 2003. — 169 p. — (Les références). — Bibliogr. Index. — ISBN 2-7081-2872-8

CROCHET DAMAIS, 2003

💻 **Panorama des outils de recherche** [en ligne] / Antoine Crochet Damais. — Journal du Net, 2 octobre 2003 [consulté le 5 octobre 2003]
< http://solutions.journaldunet.com/0310/031002_pano_moteur >

CROCHET DAMAIS, 2002

Comparez les moteurs de recherche [en ligne] / Antoine Crochet Damais. — Journal du Net, 1 janvier 2002 [consulté le 30 septembre 2003]
< http://solutions.journaldunet.com/0203/020312_moteur.shtml >

DALBIN, 2002

Indexation automatique et langage naturel [en ligne] / Sylvie Dalbin. — Rouen, ADBS, 2002 [consulté le 10 octobre 2003]
< http://www.galilo.net/normandie-adbs/actualite/ill/dalbin_rouen2002v2.pdf >

DALBIN, 2000

Une expérience d'utilisation d'un système d'information documentaire en langage naturel / Sylvie Dalbin ; Bruno Saleras. — Documentaliste - Sciences de l'information, décembre 2000, vol. 37, no. 5-6, p. 312-324

DECHILLY, 2000

Une ontologie pour éditer des schémas de description audiovisuels, extension pour l'inférence sur les descriptions [en ligne] / Thomas Dechilly ; Bruno Bachimont. Paris, INA, 2000 [consulté le 12 septembre 2003]
< http://www.irit.fr/IC2000/ACTES/DechillyIC00.pdf >

Présente l'ontologie du domaine audiovisuel, en cours de développement à l'INA, liée à la norme MPEG-7.

DÉGEZ, 2001

Thésauroglossaire des langages documentaires : un outil de contrôle sémantique / Danièle Dégez, Dominique Ménillet. — Paris : ADBS éd., 2001. — 181 p. — (Collection Sciences de l'information. Série Recherches et documents, ISSN 1159-7666). — Bibliogr. - ISBN 2-84365-051-8

DEXTRE CLARKE, 2003

BS 8723 : a new British Standard for structured vocabularies [en ligne] / Stella G. Dextre Clarke. — [consulté le 30 septembre 2003]
< http://www.glam.ac.uk./soc/research/hypermedia/NKOS-workshop%20Folder/dextre_clarke.ppt >

DING, 2002a

Ontology research and development. Part 1 : a review of ontology generation / Ding Ying ; Schubert Foo. — Journal of information science, 2002, vol. 28 no. 2, p. 123-136

DING, 2002b

Ontology research and development. Part 2 : a review of ontology mapping and evolving / Ding Ying ; Schubert Foo. — Journal of information science, 2002, vol. 28 no. 2, p. 375-388

EBSI, 2002

Terminologie de base en sciences de l'information : volets 1 et 2 [en ligne] / Université de Montréal, Faculté des arts et des sciences, École de bibliothéconomie et des sciences de l'information. — Montréal : Université de Montréal, mise à jour le 16 octobre 2002 [consulté le 27 septembre 2003]
< http://www.ebsi.umontreal.ca/termino/index.htm >

FARNUM, 2002

Redesigning an e-business taxonomy : Egreetings project case study / Chris Farnum. — Bulletin of the American Society for Information Science and Technology, June-July 2002, vol. 28, no. 5, p. 10-13

FUCHS, 1993

Linguistique et traitements automatiques des langues / par Catherine Fuchs ; avec la collab. de Anne Lacheret-Dujour et de Bernard Victorri. — [Paris] : Hachette supérieur, 1993. — 303 p. — (HU. Linguistique) (Hachette université. Langue, linguistique, communication). — Bibliogr. Index. — ISBN 2-01-016908-5

GANDON, 2002

Distributed artificial intelligence and knowledge management : ontologies and multi-agent systems for a corporate semantic web [en ligne] / Fabien Gandon. — Nice : INRIA, 2002. — Thèse de doctorat en STIC soutenue en novembre 2002 [consulté le 3 août 2003]

GILCHRIST, 2001

Corporate taxonomies : report on a survey of current practice / Alan Gilchrist. — Online Information Review, 2001, vol. 25, no. 2, p. 94-102

GIRARDEAU, 2002

Analyser un logiciel de GED : les fonctionnalités passées au crible / Nicolas Girardeau. — In : GED et solutions logicielles : guide pratique. — Paris : Archimag, 2002. — 96 p.

GRUBER, 1995

Toward Principles for the Design of Ontologies Used for Knowledge Sharing [en ligne] / Thomas R. Gruber. — International Journal of Human and Computer Studies, 1995, no. 43, p. 907-928
< http://ksl-web.stanford.edu/knowledge-sharing/papers/onto-design.rtf >

GRUBER, 1993

A Translation Approach to Portable Ontology Specifications [en ligne] / Thomas R. Gruber. — Knowledge Acquisition, 1993, no. 5, p. 199-220
< http://www.ksl.Stanford.EDU/knowledge-sharing/papers/ontolingua-intro.rtf >

HABERT, 1997

📖 **Les linguistiques de corpus** / Benoît Habert, Adeline Nazarenko, André Salem. — Paris : A. Colin, 1997. — 240 p. — (Collection U. Série Linguistique). — Bibliogr. Index. — ISBN 2-200-01775-8

HUDON, 1994

📖 **Le Thésaurus : conception, élaboration, gestion** / par Michèle Hudon. — Québec : Éd. ASTED, 1994. — 220 p. — (Clé en main). — Bibliogr. p. 212-213. Index. — ISBN 2-921548-14-3

IDC, 2001

💻 **The High Cost of Not Finding Information** [en ligne] / IDC. — [consulté le 13 octobre 2003]
< www.knowledge-wave.com/scripts-include/ en-us/downloads/idcinfo2996.pdf >

INFLUO SOFTWARE, 2002

💻 **Moteurs de recherche : où est la technologie ?** [en ligne] / Influo Software. — [consulté le 10 septembre 2003]
< http://h2ptm.hymedia.univ-paris8.fr/webmining/prez/Influo%20WP%20moteurs%20V1.pdf >

ISO/IEC, 2002

💻 **ISO-IEC 13250 Topic Maps : Information Technology : Document Description and Processing Languages** [en ligne] / ISO ; IEC. — [consulté le 5 octobre 2003]
< www.y12.doe.gov/sgml/sc34/document/ 0322_files/iso13250-2nd-ed-v2.pdf >

JACQUEMIN, 2000

📖 **Traitement automatique des langues pour la recherche d'information** / sous la dir. de Christian Jacquemin. — Paris : Hermès, 2000. — p. 327-591. — Notes bibliogr. — No de : "TAL. Traitement automatique des langues", ISSN 1248-9433, vol. 41, no. 2, 2000. — ISBN 2-7462-0225-5

JOLION, 2000

📖 **L'Indexation** / sous la dir. de Jean-Michel Jolion. — [Paris] : [Hermes sciences publications], 2000. — 182 p. — No de : "Document numérique", ISSN 1279-5127, vol. 4, no. 1-2, 2000. — ISBN 2-7462-0224-7

KAYSER, 2001

📖 **Traitement automatique du langage naturel** / sous la dir. de Daniel Kayser et Bernard Levrat. — Paris : Hermès, 2001. — p. 249-494. — Notes bibliogr. — No. de "TSI. Technique et science informatiques", ISSN 0752-4072, vol. 20, no. 3, 2001. — ISBN 2-7462-0256-5

KNOX, 2003

Taxonomy Development : Build or Buy ? / R. Knox, K. Harris, F. Caldwell, D. Logan. — Research Note, Giga Information Group, 9 September 2003

LAVIALE, 1999

La Cyberthèque de la Direction informatique de la Société Générale : un outil de gestion et de partage des connaissances / Laure Laviale. — Paris : INTD, 1999. — 91 p. — Mémoire INTD

LE COADIC, 2001

Dictionnaire encyclopédique de l'information et de la documentation / [réd. par] Yves F. Le Coadic, Michel Melot, Paul-Dominique Pomart [et al.] ; [sous la dir.] de Serge Cacaly. — [Paris] : Nathan, 2001. — 634 p. — ISBN 2-09-191252-2

LE MOAL, 2002

La recherche d'information sur les réseaux : cours INRIA, 30 septembre - 4 octobre 2002, Le Bono, Morbihan / ouvrage coordonné par Jean-Claude Le Moal, Bernard Hidoine et Lisette Calderan. — Paris : ADBS éd., 2002. — 322 p. — (Collection Sciences de l'information. Série Études et techniques, ISSN 1160-2376). — Notes bibliogr. — ISBN 2-84365-062-3

LEFÈVRE, 2000

La recherche d'informations : du texte intégral au thésaurus / Philippe Lefèvre. — Paris : Hermès science publications, 2000. — 253 p. — ISBN 2-7462-0173-9

LELOUP, 2002

Catégorisation et classification automatiques [en ligne] / Catherine Leloup. — 19 avril 2002 [consulté le 15 août 2003]
< http://www.adbs.fr/uploads/journées/leloup/index.htm >

LELOUP, 1998

Moteurs d'indexation et de recherche : environnements client-serveur, Internet et intranet / Catherine Leloup. — Paris : Eyrolles, 1998. — XIII-285 p. + 1 CD-ROM. — Bibliogr. Glossaire. Lexique français-anglais et anglais-français. Index. — ISBN 2-212-08976-7

MAHMOUDI, 1997

Traitement automatique des langues naturelles : évolution et perspectives / Seyed Mohammad Mahmoudi. — Revue de la science de l'information, juillet 1997, no. 6, p. 9-60

MANIEZ, 2002

📖 **Actualité des langages documentaires : fondements théoriques de la recherche d'information** / Jacques Maniez. — Paris : ADBS éd., 2002. — 395 p. — (Collection Sciences de l'information. Série Études et techniques, ISSN 1160-2376). — Bibliogr. p. 365-373. Index. — ISBN 2-84365-060-7

MANIEZ, 1999

📖 **Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information** / [textes réunis par] Jacques Maniez, Widad Mustafa El Hadi. — Villeneuve d'Ascq : Université de Lille 3, 1999. — 403 p. — (Travaux et recherches — Université de Lille 3). — Notes bibliogr. Index. — ISBN 2-84467-002-4

MARCHAND, 1998

📖 **L'analyse du discours assistée par ordinateur : concepts, méthodes, outils** / Pascal Marchand. — Paris : A. Colin, 1998. — 222 p. — (U. Psychologie). — Bibliogr. Notes bibliogr. Index. — ISBN 2-200-21914-8

MEMODATA, 1999

💻 **Memodata** [en ligne] / Memodata. — Caen : Memodata, 10 sept. 1999, dernier ajout octobre 2003 [consulté le 22 octobre 2003]
< http://www.memodata.com/index.shtml >

METZGER, 2001

📖 **L'indexation à l'ère d'Internet : actes du congrès d'ISKO-France, École nationale supérieure des sciences de l'information et des bibliothèques et Université Jean Moulin Lyon-3, Lyon, 21-22 octobre 1999** / éd. Jean-Paul Metzger, Mohamed Hassoun, Omar Larouk. — [Paris] : ISKO-France ; Villeurbanne : ENSSIB, 2001. — 240 p. — Notes bibliogr. — ISBN 2-910227-39-1

MINEL, 2002

📖 **Filtrage sémantique : du résumé automatique à la fouille de textes** / Jean-Luc Minel. — [Paris] : Hermès science publications ; [Cachan] : Lavoisier, 2002. — 202 p. — Bibliogr. p. 187-199. Index. — ISBN 2-7462-0602-1

MONDECA, 2002

💻 **Making sense of content** [en ligne] / Mondeca. — Paris, Mondeca, 2003 [consulté le 5 octobre 2003]
< http://www.mondeca.com/english3/published-doc/ITM-doc8p-fr.pdf >

MORGAT, 1995

📖 **Audit et gestion stratégique de l'information** / Pierre Morgat. — Paris : les Ed. d'Organisation, 1995. — 143 p. — (Collection Audit). — Bibliogr. p. 142-143. — ISBN 2-7081-1826-9

NOY, 2001

Ontology development 101 : A Guide to Creating Your First Ontology [en ligne] / Natalya F. Noy ; Deborah L. McGuinness. — Stanford, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001 [consulté le 27 juillet 2003]
< http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf >

OTIL, 2003

Lexique OTIL [en ligne] / Observatoire du traitement informatique des langues et de l'inforoute. — Liège : OTIL, dernière modification 1 août 2003 [consulté le 27 septembre 2003]
< http://www.owil.org/fr_lexique.htm >

PEPPER, 2000

The TAO of Topic Maps : Finding the Way in the Age of Infoglut [en ligne] / Steve Pepper. — [consulté le 7 octobre 2003]
< www.gca.org/papers/xmleurope2000/pdf/s11-01.pdf >

PICQ, 1999

Manager une équipe projet / Thierry Picq. — Paris : Dunod, 1999. - VI-225 p. — (Fonctions de l'entreprise. Série Animation des hommes). — La couv. porte en plus : "pilotage, enjeux, performance". - Bibliogr. — ISBN 2-10-004031-6

PIERREL, 2000

Ingénierie des langues / sous la dir. de Jean-Marie Pierrel. — Paris : Hermès science publications, 2000. — 354 p. — (Traité IC2 : information, commande, communication. Série Informatique et systèmes d'information). — Notes bibliogr. Index. — ISBN 2-7462-0113-5

RAMOS, 2003

Best practices in Taxonomy Development and Management / Laura Ramos ; Daniel W. Rasmus. — Planning Assumption, Giga Information Group, 8 January 2003

RAMOS, 2002

Taxonomy, Thesaurus, Tagging : Balancing Automation and Editorial Review / Laura Ramos. — Planning Assumption, Giga Information Group, 8 january 2003

RASTIER, 1994

Sémantique pour l'analyse : de la linguistique à l'informatique / François Rastier, Marc Cavazza, Anne Abeillé. — Paris ; Milan ; Barcelone : Masson, 1994. — XII-240 p. — (Sciences cognitives, ISSN 0991-577X). — Bibliogr. p. 225-234. Index. - ISBN 2-225-84537-9

RATH, 2003

The Topic Maps Handbook : White Paper [en ligne] / H. Holger Rath. — Gutersloh, Germany : Empolis, Arvato Knowledge Management, 2003 [consulté le 16 août 2003]
< http://www.empolis.com/download/docs/whitepapers/empolistopicmapswhitepaper_eng.pdf >

ROBERGET, 2003

Le marché impose son modèle économique à la recherche sur internet [en ligne] / Olivier Roberget. — 01 Informatique, 5 septembre 2003, no. 1735, p. 6-7 [consulté le 20 septembre 2003]
< http://www.01net.com/Pdf/01I200309051735006.pdf >

ROUMIEUX, 2000b

Linguistique : domestiquer l'ordinateur / Olivier Roumieux ; Michel Remize ; Charlotte Perrin. — Archimag, novembre 2000, no. 139, p. 21-31

SABAH, 1989

L'intelligence artificielle et le langage. 2, Processus de compréhension / Gérard Sabah. — Paris : Hermès, 1989. — 411 p. — (Langue, raisonnement, calcul, ISSN 0988-0569). — Bibliogr. p. 375-395. Index. — ISBN 2-86601-187-2

SAEED, 2002

Using Dewey decimal classification scheme (DDC) for building taxonomies for knowledge organisation / Hamid Saeed ; Abdus Sattar Chaudhry. — Journal of Documentation, 2002, vol. 58, no. 5, p. 575-583

SATO, 2001

Glossaire des termes d'ATO [en ligne] / Université du Québec à Montréal, Service ATO ; École de bibliothéconomie et des sciences de l'information ; Documensa. — Montréal : Université du Québec à Montréal, [consulté le 27 septembre 2003]
< http://www.ling.uqam.ca/sato/glossaire/index.html >

SAVOY, 1993

Stemming of french words based on grammatical categories / Jacques Savoy. — Journal of ASIS, January 1993, vol. 44, no. 1, p.1-9

SILBERZTEIN, 1993

Dictionnaires électroniques et analyse automatique de textes : le système INTEX / Max Silberztein. — Paris ; Milan ; Barcelone : Masson, 1993. — 233 p. — (Informatique linguistique). — Bibliogr. p. 187-191. Index. — ISBN 2-225-84157-8

SMITH, 2003

Concept-based Learning Spaces : Apply domain-specific KOS principles for organizing collections/services for given applications [en ligne] / Terence M. Smith ; Marcia Zeng. — Kent, Ohio : Network Knowledge Organization Systems, last updated 16 may 2003 [consulté le 27 septembre 2003]
< http://nkos.slis.kent.edu/2003workshop/zengnkos.ppt >

SOERGEL, 2003

From legacy KOS to full-fledged ontologies [en ligne] / Dagobert Soergel ; Kathy Newton. — Kent, Ohio : Network Knowledge Organization Systems, last updated 16 may 2003 [consulté le 27 septembre 2003]
< http://nkos.slis.kent.edu/2003workshop/soergel.ppt >

TRIPLEHOP TECHNOLOGIES, 2002

Match Point, Information Retrieval Software : Differentiating Factors [en ligne] / TripleHop Technologies. — [consulté le 16 septembre 2003]
< www.triplehop.com/pdf/ MatchPoint%20Competitive%20Advantage%20White%20Paper.pdf >

VACARIE, 2001

Étude comparative des moteurs de recherche sémantique : choix d'un outil pour le service de veille du département informatique de la Société Générale / Cécile Vacarie. — Paris : INTD, 2001. — 135 p. — Mémoire INTD

VERITY, 2003

The Verity K2 Discovery Tier : the Importance of Advanced, Effective Search Tools : Verity White Paper [en ligne] / Verity. — Sunnyvale, California : Verity, 2003 [consulté le 16 septembre 2003]
< www.verity.co.uk/pdf/white_papers/ MK0348a_Discovery_WP.pdf >

VERITY, 2002

The ABCs of Content Organization : Verity White Paper [en ligne] / Verity. — Sunnyvale, California : Verity, 2002 [consulté le 16 septembre 2003]
< www.verity.com/pdf/white_papers/ MK0391a_ContentOrg_WP.pdf >

VERITY, 2001a

Verity Collection Reference Guide V4.0 / Verity. — Sunnyvale, California : Verity, 2001

VERITY, 2001b

Verity Internationalization : Enabling E-business in Multiple Languages : Verity White Paper [en ligne] / Verity. — Sunnyvale, California : Verity, 2001 [consulté le 16 septembre 2003]
< www.verity.com/company/contact/international/ pdf/MK0375b_I18N_WP.pdf >

VERITY, 2001c

Verity K2 Enterprise, Fundamentals Guide V4.0 / Verity. — Sunnyvale, California : Verity, 2001

VERITY, 2001d

Verity K2 Enterprise, Intelligent Classification Guide V4.0 / Verity. — Sunnyvale, California : Verity, 2001

VERITY, 2001e

Verity K2 Search Objects Guide V4.0 / Verity. — Sunnyvale, California : Verity, 2001

VERITY, 2001f

Verity Query Language Guide V4.0 / Verity. — Sunnyvale, California : Verity, 2001

VICKERY, 1997

Ontologies / B. C. Vickery. — Journal of information science, 1997, vol. 23, no. 4, p. 277-286

VILLACAMPA, 1999

La création d'une base de donnés avec Retrievalware d'Excalibur / Alberto Villacampa. — Paris : INTD, 1999. — 43 p. — Mémoire INTD

VILLOING, 2003

Influo Software : serveurs de connaissances, la neuromimétique au service d'un Web sémantique [en ligne] / F. Villoing. — 14 ianvier 2003 [consulté le 15 septembre 2003]
< http://h2ptm.hymedia.univ-paris8.fr/webmining/prez/influo_fichiers/frame.htm >

WARNER, 2003

Guidelines and Principles for Developing Search and Browse Vocabularies [en ligne] / Amy J. Warner. — Kent, Ohio : Network Knowledge Organization Systems, last updated 16 may 2003 [consulté le 27 septembre 2003]
< nkos.slis.kent.edu/2003workshop/amy.ppt >

ZWEIGENBAUM, 1996

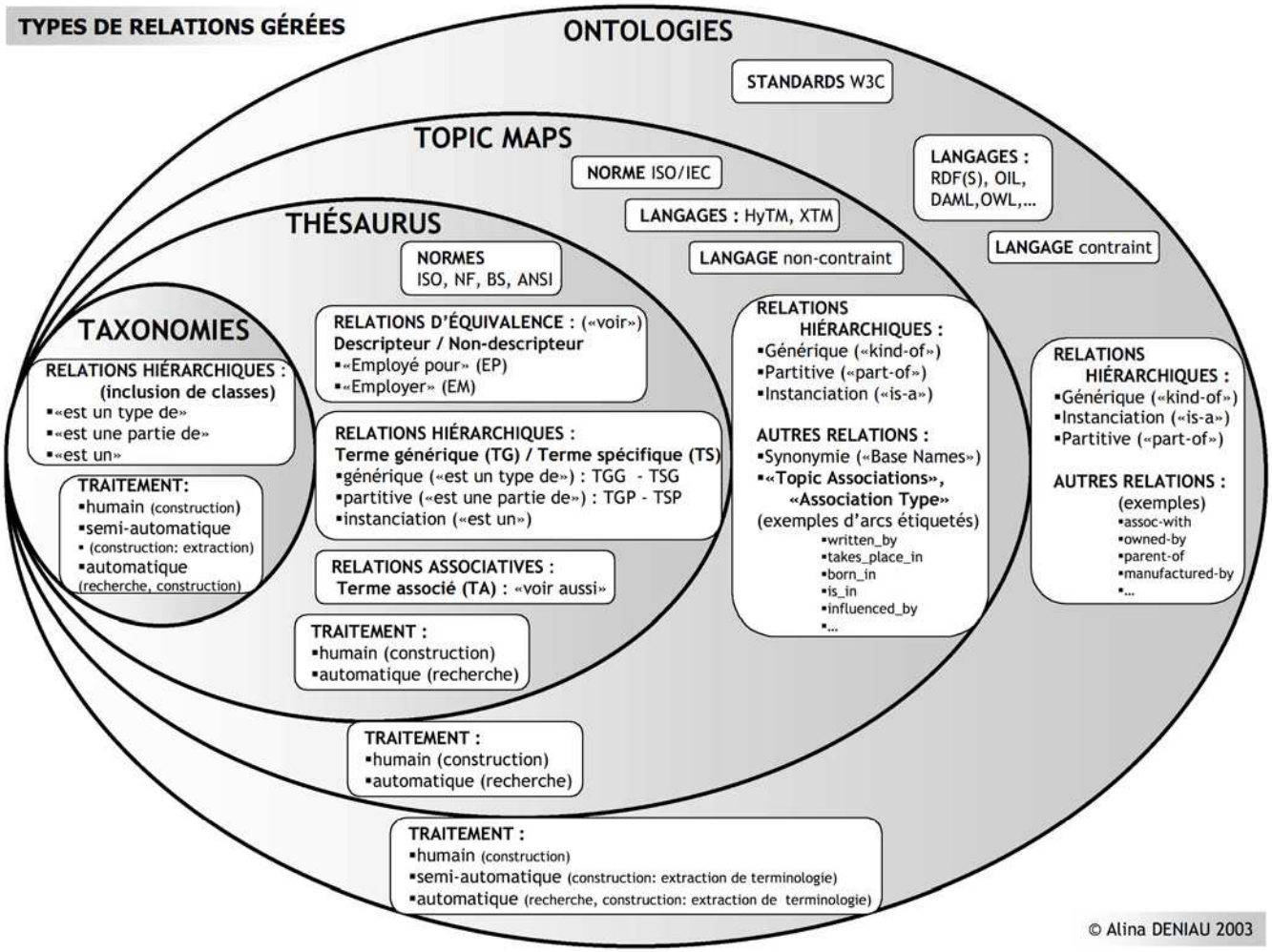
Le rôle du lexique sémantique et de l'ontologie dans le traitement automatique de la langue médicale [en ligne] / Pierre Zweigenbaum ; Bruno Bachimont ; Jacques Bouaud,... [et al.]. — [consulté le 25 août 2003]
< http://www-test.biomath.jussieu.fr/~pz/FTPapiers/Zweigenbaum:CRISTALS96.pdf >

ANNEXES

SCHÉMA TAXONOMIE / THÉSAURUS / TOPIC MAPS / ONTOLOGIE

Le schéma se trouve page suivante.

Représentation schématique des types de relations gérées par les systèmes d'organisation des connaissances (KOS)



OPÉRATEURS VERITY

TABLEAU DES OPÉRATEURS UTILISÉS PAR VERITY (VERITY QUERY LANGUAGE - VQL)

(dans l'application Cyberthèque)

En gras : les opérateurs utilisés actuellement
En italique : les opérateurs qui améliorent les résultats actuels

FRANÇAIS	ANGLAIS	ACTION	OBSERVATIONS
Evidence Operators	**Evidence Operators**	Classe d'opérateurs à utiliser au niveau des chaînes de caractères (Evidence Level)	
Mot	**Word**	Sélectionne les documents contenant une ou plusieurs occurrences du mot tel quel.	
Racine	**Stem**	Lemmatisation : sélectionne le mot et ses variantes grammaticales (ex. : pour les noms : pluriel, féminin ; pour les verbes : déclinaisons). Uniquement le lemmatiseur français a été installé.	ATTENTION ! : Ne jamais utiliser pour les sigles. ATTENTION ! : Le lemmatiseur anglais n'est pas installé : ne jamais demander la lemmatisation des mots anglais. (Utiliser l'opérateur <Mot> à la place.)
Consonnance	Soundex	Sélectionne les occurrences du terme et les mots qui « sonnent comme » celui-ci.	ATTENTION ! : Un index des variantes doit préalablement être construit dans les collections. (ex. : liste des graphies possibles du nom de Khadafi ou de Mao Tze Toung...)
Synonyme	Thesaurus	Sélectionne le terme et ses synonymes présents dans le dictionnaire standard.	ATTENTION ! : Source de bruit dans un corpus métier. Utiliser uniquement sur des corpus généralistes (ex. : moteur mis en place dans un site Web).

Typo	TypoEqual	Sélectionne les mots "similaires". Des algorithmes calculent la distance possible entre deux caractères (ex.: n et h).	ATTENTION ! : Source de bruit. Utiliser uniquement sur des textes obtenus par scan + OCR.
Troncature	Wildcard	SYMBOLES UTILISES : ? (= 1 caractère) * (= 1 ou plusieurs caractères) [ ] (= 0, 1 ou plusieurs caractères) { } (= n'importe lequel des caractères spécifiés à l'intérieur) ^ (= dans une parenthèse, exclut les caractères spécifiés après) - (= dans une parenthèse, sélectionne les caractères situés entre les caractères spécifiés)	L'astérisque (*) peut être utilisé comme troncature droite ou gauche. ATTENTION ! : L'usage des parentèses en interrogation demande un certain apprentissage.
Combinatorial Operators	**Combinatorial Operators**	Classe d'opérateurs à utiliser aux niveaux 1 et 2 (Topics et Sub-Topics)	
Cumul	**Accrue**	Sélectionne les documents qui contiennent au moins un des termes.	ATTENTION : ! Il s'agit d'un OU pondéré, au comportement différent du OU qu'on utilise sur le web. Pondération non-linéaire.
Ou	**Or**	Sélectionne les documents qui contiennent au moins un des termes.	ATTENTION : ! Il s'agit d'un OU pondéré, au comportement différent du OU qu'on utilise sur le web. Pondération particulière : le score le plus haut est adopté par le noeud.
Et	**And**	Sélectionne les documents qui contiennent tous les termes.	ATTENTION : ! Il s'agit d'un ET pondéré, au comportement différent du ET qu'on utilise sur le web.
Quelconque	Any	Sélectionne les documents qui contiennent tous les termes.	Opérateur booléen. Equivalent du OU qu'on utilise sur le web. Pas de pondération. (Score : 0/1)
All (=Tous)	All	Sélectionne uniquement les documents qui contiennent tous les mots.	Opérateur booléen. Equivalent du ET qu'on utilise sur le web. Pas de pondération. (Score : 0/1)

Behavior Operators	**Behavior Operators**	Classe d'opérateurs qui modifient le comportement des autres opérateurs	
Plusieurs	**Many**	Modifie le score (pertinence) en fonction du nombre d'occurrences dans le document des mots (concepts ou chaînes de caractères) définis par les noeuds inférieurs. Plus un document contient d'occurrences d'un concept, plus son score augmentera.	
Casse	*Case*	Rend sensible à la casse des chaînes de caractères. (Par défaut le moteur n'est pas sensible à la casse.)	ATTENTION ! : A utiliser uniquement au niveau des chaînes de caractères (Evidence level). EXEMPLE : XML et xml donnent le même résultat.
Sauf	*Not*	Exclue tous les documents qui contiennent ce terme.	Opérateur booléen. Equivalent du " - " sur le web.
Ordre	*Order*	Sélectionne uniquement les documents qui contiennent les mots dans l'ordre indiqué (du haut vers le bas des noeuds subordonnés).	
Opérateurs de proximité	**Proximity Operators**	Classe des opérateurs de proximité	ATTENTION ! : L'option PSW doit avoir été utilisée lors de la création de la collection
Proche (Proche/n)	**Near (Near/n)**	Sélectionne les documents qui contiennent tous les termes. Proche est pondéré par un algorithme qui calcule la proximité relative des termes. Proche/n permet de spécifier la distance entre les 2 mots).	
Expression	**Phrase**	Sélectionne les documents qui contiennent tous les termes dans la même expression, dans l'ordre (du haut vers le bas du concept).	
Phrase	Sentence	Sélectionne les documents qui contiennent tous les termes dans la même phrase.	ATTENTION ! : Par défaut, n'est pas sensible à l'ordre des mots. Pour le forcer, utiliser <Ordre>.
Paragraphe	Paragraph	Sélectionne les documents qui contiennent tous les termes dans le même paragraphe.	

Opérateurs de sélection de champ et zone	**Field and zone operators**	Classe d'opérateurs agissant uniquement sur des documents structurés	Inutiles sur du non-structuré, ils ne peuvent pas agir sur le web. (ex. : pour le spider)
Champ	Field	Sélectionne le champ (ex.date ou source) dans lequel effectuer la recherche.	ATTENTION ! : Il faut connaître le nom exact du champ à sélectionner (ex. si la base contient plusieurs champs date).
Dans	In	Sélectionne une zone de recherche dans un document.	ATTENTION ! : Il faut avoir sélectionné cette option au moment de la constitution de la collection.
Sur les champs on peut utiliser les opérateurs suivants :			Non utilisés
Contenu	Contains		
Correspondance	Matches		
Début	Starts		
Fin	Ends		
SousChaîne	Substring		
... et les opérateurs de comparaison numérique suivants :			Non utilisés
<			
<=			
=			
>=			
>			
Score Operators	**Score Operators**	Interviennent dans le calcul des pondérations	Non utilisés
	Complement		
	LogSum/n		
	Mult/n		
	Product		
	Sum		
	YesNo		

NOTICE :

Moteurs de recherche et restitution de l'information dans les grandes entreprises : l'exemple du portail Cyberthèque de la Direction des Systèmes d'Information de la Société Générale / Alina Ivanciuc Deniau. — Paris : INTD, 2003. — 154 p. — Bibliogr. p. 112-146. — Mémoire INTD

RÉSUMÉ :

Après avoir replacé les moteurs de recherche dans le contexte de la recherche d'information et des langages documentaires (notamment le thésaurus), compte tenu des mutations de ces dernières années (taxonomies, ontologies, Topic Maps), la première partie du mémoire se propose de décrire le fonctionnement de ces outils issus de la recherche en traitement automatique du langage (TALN). La définition du TALN, en soulignant les apports de chaque discipline impliquée avec un éclairage particulier sur la linguistique, est suivie d'une typologie des produits présents dans les grandes entreprises. Le fonctionnement des moteurs de recherche est décrit ensuite à travers les opérations effectuées par les moteurs de recherche linguistiques pour traiter la masse d'information textuelle lors de l'indexation. Cette description prend pour exemple un produit particulier : « K2 Enterprise » de la société Verity. La seconde partie retrace l'audit effectué afin de mettre en place des améliorations dans un portail d'entreprise dont la recherche est gérée par le moteur de recherche Verity K2 : la Cyberthèque de la Direction des Systèmes d'Information de la branche Banque de Détail de la Société Générale, portail de veille technologique et concurrentielle.

MOTS-CLÉS :

MOTEUR DE RECHERCHE ; VERITY ; INDEXATION AUTOMATIQUE ; LANGAGE DOCUMENTAIRE ; THÉSAURUS ; TAXONOMIE ; TOPIC MAPS ; ONTOLOGIE ; TALN ; ANALYSE LINGUISTIQUE ; SEGMENTATION ; LEMMATISATION ; ÉTIQUETAGE ; CATÉGORISATION AUTOMATIQUE ; CLASSIFICATION AUTOMATIQUE ; RÉSUMÉ AUTOMATIQUE