THÉSAURUS ET INFORMATIQUE DOCUMENTAIRES

Partenaires de toujours ?

Sylvie Dalbin

A.D.B.S. | *Documentaliste-Sciences de l'Information*

2007/1 - Vol. 44
pages 42 à 55

ISSN 0012-4508

En étudiant d'abord les usages et fonctionnalités logicielles, puis les différentes familles de logiciels ou d'applicatifs, Sylvie Dalbin rappelle ici la place et le rôle souvent occultés du thésaurus dans divers logiciels documentaires pour l'indexation, la recherche et la maintenance terminologique. Si les outils assurant la gestion et l'exploitation de thésaurus sont anciens, les applications récentes révèlent une situation plus variée, avec trois orientations majeures : une distinction nette entre systèmes d'indexation et d'interrogation, un renforcement du suivi et de la maintenance des vocabulaires contrôlés et un développement d'activités liées à la conception de vocabulaires ou d'interfaces d'accès à l'information.

par **SYLVIE DALBIN**

Thésaurus et informatique documentaires
Partenaires de toujours ?

OCCULTÉES PAR LES LOGICIELS d'indexation et de recherche automatiques (moteurs de recherche ou logiciels en langage naturel) ou subordonnées aux logiciels de gestion et recherche documentaires (LGRD), les applications logicielles permettant d'exploiter les thésaurus documentaires[1], de les développer et de les maintenir ne font que très rarement l'objet d'études et d'articles techniques et professionnels.

Il existe, dans les pays anglo-saxons[2], une riche production éditoriale, déjà ancienne, complétée par une veille importante sur une catégorie particulière de logiciels dédiés aux thésaurus : les logiciels de conception et maintenance de thésaurus dits indépendants ou autonomes (*standalone*) [4] [5] [7] [8], largement inusités en France[3]. Quant aux modules « thésaurus » intégrés aux logiciels de gestion et recherche documentaires (LGRD) et utilisés par les professionnels de l'information-documentation, ils sont abordés, mais de façon succincte[4], dans les études francophones portant sur ces LGRD. Nous pouvons toutefois mentionner deux études, anciennes, qui présentent les

fonctionnalités et caractéristiques de ces modules « thésaurus » [1] [6].

Il faut aussi noter que les spécifications fonctionnelles dédiées aux thésaurus, rarement détaillées dans les cahiers des charges lors de l'acquisition d'une solution documentaire, ne sont concrètement pas utilisées comme critères de sélection d'un LGRD. Les difficultés surviennent donc après coup, au moment de passer aux activités de production documentaire. Les professionnels se rendent alors compte qu'il n'est pas possible de charger le thésaurus en l'état, que le nouveau module ne permet pas de gérer au sein du thésaurus un champ « Candidat descripteurs » pourtant indispensable à la maintenance de ce thésaurus, que la recherche d'un descripteur par les synonymes n'est pas possible dans les activités d'indexation, ou encore que le thésaurus n'est pas du tout consultable à la recherche… Le terme même régulièrement employé de *gestion informatisée de thésaurus* ne donne pas la mesure des spécifications fonctionnelles utiles à une réelle exploitation (et non gestion) des thésaurus pour l'indexation et pour la recherche.

Et pourtant les thésaurus n'ont jamais fait autant l'objet d'attention[5] dans des lieux variés et par des spécialistes de toutes origines : ils sont intégrés à des publications dédiées à des spécialistes des « technologies de l'information » pour les intranets[6], supports du développement du web sémantique en tant que « systèmes simples d'organisation de la connaissance[7] » ou encore étudiés de près par les concepteurs d'interfaces web. Toutes ces investigations ouvrent des perspectives certes différentes des solutions traditionnelles, mais toujours riches et passionnantes.

Nous nous proposons dans cet article de faire un point de la situation en France de l'informatique associée aux thésaurus documentaires sur le plan des fonctionnalités (section 1) puis plus brièvement sur celui des familles de logiciels ou d'applicatifs (section 2). Pour mieux situer cette contribution, nous commencerons par un rapide exposé du contexte dans lequel se place le thésaurus.

Sylvie Dalbin
est consultante en organisation et ingénierie documentaires depuis 1989 au sein d'Assistance & Techniques Documentaires. Dès 1986, alors documentaliste à EDF, elle travaillait sur les questions d'indexation automatique et de recherche sur les contenus. Ses interventions dans les entreprises et les organisations portent aujourd'hui plus spécifiquement sur l'évaluation, les méthodes et les outils d'accès à l'information.
sylvieATD@aol.com
www.ATD-doc.com

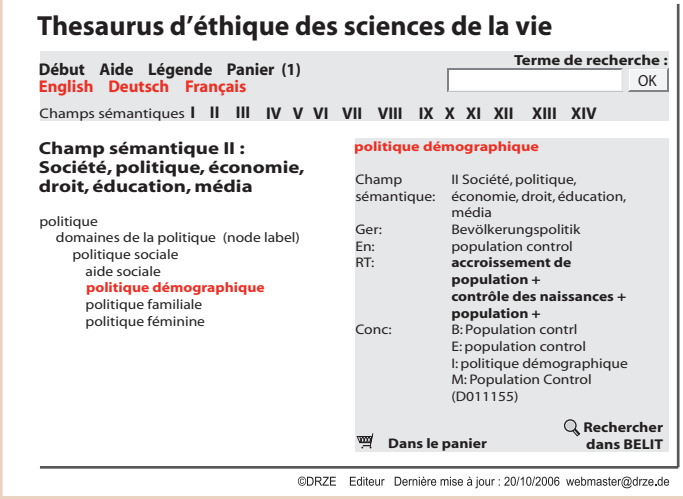
La place du thésaurus dans les systèmes d'information documentaire

Depuis le tournant du siècle, plusieurs éléments de notre environnement, étudiés jusque-là d'un peu loin par les professionnels de l'information-documentation (I-D) semblent devoir être intégrés de façon plus nette dans les projets documentaires.

• *La grogne des utilisateurs d'information face à la complexité et la multiplicité des interfaces et des systèmes* : ils souhaitent être informés « simplement », rapidement. Aujourd'hui en masse, ces utilisateurs ne sont ni des professionnels de la recherche documentaire, ni des chercheurs ou des scientifiques de laboratoire pour qui le modèle documentaire des années cinquante – une formulation claire et précise de la question – avait été initialement conçu. Il est clair que les thésaurus ou plutôt les interfaces d'interrogation actuelles ne lui sont pas réellement destinées [2]. La réponse apportée depuis dix ou quinze ans par les professionnels de l'I-D, parallèlement aux fonctions de *push*[8], repose sur la recherche dite *plein texte* sur l'ensemble des notices sans réelle évaluation de l'efficacité d'un tel montage et surtout sans lui adjoindre les outils (filtres, pondération, expansion) adaptés à ce type de fonctionnement ; des outils que les serveurs des banques de données professionnelles proposaient dès les années quatre-vingt, avec les techniques informatiques de ►

1 Nous employons le terme de *thésaurus documentaires* pour les distinguer des thésaurus de langue, comme le *Roget's Thesaurus* ou le *Thésaurus Larousse*. Autre formulation employée : *thésaurus de descripteurs*, par Georges Van Slype (*Conception et gestion des systèmes documentaires*, Les Éditions d'Organisation, 1987, page 89) ou Michèle Hudon (*Le thésaurus : conception, élaboration, gestion*, ASTED, 1994, page 35).
2 De nombreuses ressources allemandes disponibles sur le Web montrent que ces outils y sont également utilisés et étudiés. Mentionnons le logiciel SuperThes pour le thésaurus GEMET, mais aussi les produits IC INDEX 5.0, MIDOSThesaurus ou THESMain/THESshow, bien souvent associés à une problématique de terminologie et de traitement multilingue comme l'indiquent les présentations des produits (*mehrsprachigen / multilingualer*).
3 Une pré-enquête sur les logiciels utilisés avec les thésaurus, conduite entre septembre et octobre 2006 auprès d'une trentaine d'adhérents de l'ADBS membres de groupes sectoriels, a permis de valider cette hypothèse : sur 27 répondants, quatre seulement indiquaient avoir connaissance de ce type de logiciels, une seule personne ayant eu l'occasion d'en utiliser.
4 Voir les publications du cabinet Tosca Consultants aux éditions de l'ADBS ou les Guides pratiques édités par le magazine *Archimag*. Le constat est le même pour le suivi des logiciels libres (Camille Espiau, *Projet SIGB Libres : étude comparative des fonctionnalités des SIGB libres*, mise à jour novembre 2006, sur le site collaboratif www.sigb-libres.info/).
5 Nous n'abordons pas ici les questions du coût de l'indexation humaine ou du développement des techniques d'indexation automatiques ou de recherche linguistique. Voir *infra*, ainsi que : Jacques Chaumier et Martine Dejean, « Recherche et analyse de l'information textuelle : tendances des outils linguistiques », *Documentaliste – Sciences de l'information*, 2003, vol. 40, n° 1, p. 14-24.
6 Un chapitre entier est consacré aux vocabulaires contrôlés dans l'ouvrage de Peter Morville et Louis Rosenfeld, *Information Architecture for the World Wide Web*, 2nd ed., 456 pages, O'Reilly, décembre 2006.
7 Voir SKOS, page 75.
8 DSI, diffusion sur profil, puis RSS.
9 Voir « Thésaurus et informatique documentaires : des Noces d'Or », pages 76 et suivantes.

FIGURE 1 – THESAURUS D'ETHIQUE DES SCIENCES DE LA VIE



www.drze.de/BELIT/thesaurus/show_tree.html?nr=22&la=fr

l'époque[9]. Le résultat n'est pas toujours à la hauteur des espérances.

• La multiplication des ressources et des flux d'information et de documents interdit maintenant d'imaginer la constitution d'une base homogène sur le plan des traitements documentaires. *L'objectif de la base documentaire unique semble révolu*. Il est remplacé par le modèle de l'accès simplifié et unifié à des ressources multiples, dont certaines sont traitées dans d'autres lieux et avec d'autres outils linguistiques.

• L'utilisateur veut être informé, ce qui ne veut pas dire uniquement obtenir un lot de références, mais plutôt les informations utiles contenues dans le document. Ce phénomène s'accentue avec l'évolution parallèle de pratiques différentes de production de documents numériques. De plus, ces documents utiles aux usagers sont disponibles sous format numérique, dans un pourcentage variable, bien sûr, suivant les environnements professionnels et les activités des utilisateurs. Dans ce contexte de production d'information numérique, *la notice perd son statut central face à un document multiforme contenant ses propres métadonnées ou articulé avec elles*.

• C'est dans ce contexte que *le couplage des logiciels documentaires traditionnels avec des moteurs linguistiques s'installe dans nos environnements professionnels*, y compris dans des lieux où la présence du document numérique est faible[10], ce qui démontre que l'accès à l'information peut suivre des filières distinctes de celle de l'indexation de cette information.

Dans ces conditions, quelle place réserver aux vocabulaires contrôlés, et plus particulièrement aux thésaurus ? Deux situations de travail distinctes sur le plan des besoins fonctionnels autour des thésaurus nous semblent se dessiner.

Ou bien l'on travaille *au plus près des producteurs de documents*[11]. Les logiques actuelles propres à l'édition électronique – éditoriale ou plus bureautique – poussent à une structuration et un formalisme plus poussés des documents dès leur rédaction. Un nombre important de métadonnées sont maintenant produites à la source ; elles accompagnent les documents numériques. Ce phénomène s'accentue ; il permet ainsi de récupérer les métadonnées avec les documents et de les exploiter au sein des systèmes documentaires. Dans ce cadre émergent des besoins importants et autonomes de développement et de maintenance de vocabulaires contrôlés (nomenclatures, thésaurus, taxonomies). L'activité est ici focalisée sur les référentiels terminologiques et sémantiques, transversalement à leurs différents usages au sein de l'organisme. Sans être totalement autonome par rapport au « poste de l'indexeur », cette activité se situe en marge de celui-ci, avec des contraintes spécifiques.

Ou bien l'on travaille *au plus près des utilisateurs d'informations* : la problématique est alors d'offrir des accès à une très grande variété de ressources documentaires, traitées avec des outils linguistiques et selon des pratiques diverses puisque seule une partie de ces fonds est alors indexée[12]. Sans préjuger de l'architecture des données, des choix faits sur l'interface d'accès et de la prise en compte (ou non) de techniques automatiques, émerge ici la notion de *langage de recherche*[13] proposant un accès cohérent, transparent et autonome par rapport aux différents langages sources.

Dans ces deux situations, que certains trouveront extrêmes mais qui se vérifient tous les jours sur le terrain, « la » base documentaire n'est plus l'alpha et l'oméga des professionnels de l'I-D. Ce décentrage par rapport au système de production documentaire traditionnel (indexation) fait jaillir de nouveaux besoins fonctionnels pour les thésaurus (visualisation ou exploitation automatique, interopérabilité de langages, etc.) qui, s'ils ne sont pas vraiment nouveaux, sont difficiles à identifier et à prendre en compte dans la pratique professionnelle actuelle.

1 Usages et fonctionnalités logicielles

Si les caractéristiques fonctionnelles du thésaurus sont connues à travers les normes et les publications pédagogiques ou techniques qui circulent depuis quarante ans, leur forme concrète, leurs usages réels et les outils logiciels qui les supportent le sont beaucoup moins. Nous proposons dans le tableau A pages suivantes un panorama des fonctionnalités informatiques possibles et nécessaires pour assurer l'ensemble des activités tournant autour du thésaurus : indexation et recherche, développement et maintenance, évaluation. Cette grille doit permettre d'évaluer des logiciels, des modules ou des interfaces spécialisées. Certaines fonctionnalités parmi les plus problématiques, novatrices ou caractéristiques des évolutions à venir sont reprises rapidement ci-après.

Indexation et recherche

L'étude des tâches à réaliser à l'indexation et à la recherche nous conduit à préciser les fonctionnalités requises pour l'une et pour l'autre, et surtout à les distinguer [*voir ci-contre*].

▼ Langage normatif à l'indexation

Que faut-il à l'indexeur ? Le texte analysé lui offre les points d'entrée terminologiques de l'auteur. L'accès au thésaurus peut donc être rapide si le logiciel fait des propositions, par exemple à partir d'un index permuté incluant les non-descripteurs. Si le terme est un équivalent, le système place l'indexeur sur le terme préférentiel du thésaurus. Une fois positionné sur ce nœud du réseau sémantique, l'indexeur doit pouvoir rapidement visualiser la grappe sémantique complète du descripteur. S'il doute et qu'il lui est nécessaire de contrôler l'usage du terme à l'indexation, l'indexeur peut consulter soit les ►

10 Voir, par exemple, le portail Catalog+ de la bibliothèque municipale de Lyon conçu autour du produit Autonomy.
11 Cette remarque s'applique aussi bien au monde du texte qu'à celui de l'image.
12 Enquête auprès des groupes sectoriels de l'ADBS : sur 26 répondants à notre pré-enquête dont le questionnaire est exploitable, 13 seulement indexent des documents, alors que 21 interrogent des bases documentaires.
13 Voir pages 89-92 l'exemple d'OTAREN.
14 Sur ce point, il faut noter que l'évolution vers une structuration de plus en plus forte des métadonnées des ressources numériques devrait nous conduire à des modifications importantes dans nos pratiques.
15 À moins qu'il ne s'agisse de retrouver un document dont on connaît déjà l'existence, voire des éléments précis sur celui-ci. Dans ce cas de « retrouvage », il n'est pas nécessaire de déployer des traitements sophistiqués.

Indexation et recherche : deux tâches distinctes

À *l'indexation*, la recherche d'un terme dans un thésaurus est encadrée par le document en cours d'analyse, ainsi que par les autres documents déjà indexés par l'indexeur et/ou disponibles dans le fonds documentaire. La caractéristique principale de ce travail est la rigueur du choix des termes d'indexation par rapport au texte. La tâche à mener par l'indexeur est précise : bien comprendre le territoire du vocabulaire et repérer, puis choisir le(s) descripteur(s) le(s) plus adéquat(s) par rapport aux notions/sujets sélectionnés dans le document. Cette tâche est à réaliser indépendamment pour chacun des éléments de données (champs)[14] auquel est associé un vocabulaire contrôlé.
Il faut ici envisager le thésaurus comme un dictionnaire, utilisé pour contrôler ou affiner le choix d'un mot ou d'une expression. C'est cet usage qui a donné son nom au thésaurus. Or le temps passé à cette fouille est bien souvent considéré comme du temps perdu alors qu'il constitue un élément clé de la qualité de l'indexation. En fonction de la régularité et de la variabilité thématique ou sectorielle de cette activité d'indexation, cette prise de connaissance du thésaurus et des notions qu'il couvre sera plus ou moins exigeante en temps. Il est donc important d'une part d'affiner en permanence sa connaissance du thésaurus, d'autre part de ne jamais hésiter à le compulser.
Par contre, au cours d'une activité de *recherche d'information*, le contexte de l'utilisateur est plus flou[15], le système référent sur lequel il s'appuie est de fait éloigné de ceux portés par les différents dispositifs documentaires qu'il est amené à interroger. S'il est étudiant ou élève et qu'il débute une formation dans le domaine sur lequel porte sa recherche d'information, sa tâche sera doublement complexe. En quête d'information, l'utilisateur aimerait également interroger en une seule fois les différents fonds susceptibles de lui fournir des informations.
Il est évident que, dans nos dispositifs documentaires actuels, l'utilisateur final qui n'indexe pas et qui n'a pas construit le thésaurus, non professionnel de la documentation et parfois non spécialiste du domaine, affronte en réalité une tâche beaucoup plus complexe qu'un documentaliste ! Et ce n'est pas une sensibilisation d'une heure qui fera son affaire...

Tableau A - FONCTIONNALITÉS DES OUTILS INFORMATIQUES PERMETTANT L'EXPLOITATION DE THÉSAURUS

Module d'un progiciel (M) / logiciel spécialisé (LS)	
Désignation et description du vocabulaire	
Type de vocabulaire contrôlé pris en compte	
Thésaurus, taxonomie/classification, liste de synonymes et d'acronymes, lexique	
Terminologies (norme ISO 12620)	#
Nombre de vocabulaires gérés	
Normes ou standards pris en compte	
Modèle de données (description) et encodage de données (ISO10646/Unicode)	
Termes et attributs	
Identifiant, unique par terme (quel que soit son statut)	
Distinction concept / terme	
Statut des termes : candidat - descripteur - descripteur obsolète (historique) - non-descripteur - relais	#
Construction autour d'un terme préférentiel ou non	
Intégration de catégories (domaines, thèmes ou microthésaurus)	#
Gestion d'une codification	#
Prise en compte de groupe de mots comme descripteur	
Longueur maximum des termes	#
Longueur minimum des termes	
Casse préservée - Casse signifiant	#
Caractères (symbole-espace) interdits. S'il en est, le(s)quel(s)	#
Possibilité d'intégrer un objet image	#
Typage de note : d'application, historique	#
Longueur des notes	#
Autre type de note possible (définition, par exemple)	#
Origine du terme (autre type de note)	
Nombre total de termes par langage (tout statut confondu ou par statut)	
Personnalisation (structures, attributs)	#
Relations	
Type de relations : synonymie, hiérarchie, voisinage	#
Appartenance à une ou plusieurs (au choix) catégories	#
Relation combinatoire : EM: terme X ET terme Y ; EM: terme X OU terme Y	#
Hiérarchie	
Typage des relations partitives et d'instanciation	
Relation générique multiple (polyhiérarchie) - Choix entre mono- et polyhiérarchie	#
Indicateur du terme de tête (générique de plus haut niveau = TT)	#
Nombre de niveaux hiérarchiques	#
Indication des termes orphelins (sans relation)	#
Appartenance possible d'un terme candidat à une catégorie	
Equivalence	
Renvoi d'un non-descripteur à plusieurs descripteurs (Voir Relation combinatoire)	#
Typage des relations d'équivalence : synonymie, variante lexicale, abréviation, quasi-synonymie	
Version linguistique en propre pour les non-descripteurs	
Personnalisation	
Autres types de relations	#
Choix de la réciprocité des relations, en particulier associative	
Multilinguisme	
Nombre de version linguistique prise en compte	
Gestion indépendante de la langue source et de la langue cible	
Possibilité de gérer des non-descripteurs par version linguistique	
Création, gestion et contrôle	
Création	
Saisie directe, d'un ou de plusieurs termes	
Création simultanée des termes et de leurs relations	
Import de fichiers	#
* format des données (contenu) à utiliser pour l'import	
* format informatique d'import (XML, txt, csv)	#
Possibilité de gérer des extensions au vocabulaire pour des publics identifiés	
Modification	
Modification de la forme du terme	
Traitement par lots. Décrochage de branches	
Ajout d'un terme générique (modification de la hiérarchie)	
Transformation d'un non-descripteur en descripteur ou vice versa	
Possibilité de modifier des termes ou leur relation, sans avoir à les supprimer	
Suppression	
Suppression d'un terme ayant des relations. Quelles modalités :	#
• impossible avant suppression des liens	#
• raccrochement automatique des spécifiques au générique immédiatement supérieur	#
• transfert en candidat	
• autre.	#
Contrôles	
Unicité de termes (=doublons)	#
Fonction des termes (un terme ne peut pas être non-descripteur et associé à un descripteur)	#
Termes orphelins	#
Réciprocité des relations - Cohérence de la hiérarchie	#
Signalement des erreurs : trace de données en erreur éditable	#
Terme signalé à l'écran	
Fusion	
Fusion de thésaurus (quelles contraintes ?)	
Synchronisation - autonomie	
Tris	
Tri automatique des termes spécifiques ou équivalents	#

Possibilité, de conserver un ordre non alphabétique (ou tri par code et non par libellé)

Visualisation, édition et export

Sélection du support : écran, impression, export #

Sélection des critères #

Type d'édition/visualisation

- Alphabétique globale #
- Alphabétique globale avec environnement sémantique #
- Hiérarchisée (à partir des termes de têtes) #
- Systématique (à partir des domaines), avec descripteurs et non-descripteurs #
- Par version linguistique, et multilingue
- Indicateurs de facettes #
- Equivalent, avec le terme associé
- Par code
- Permutée (kwic ou kwoc) #
- Gestion personnalisée par langue des mots « vides » pour l'index permuté
- Description complète des termes #
- Liste des orphelins #
- Edition/visualisation totale, partielle, paramétrable, associée au choix du contenu
- Représentation graphique des relations
- Choix de la dénomination des relations ou type de termes

Format d'export

- Formats (contenu) : zThes, SKOS, autre DTD ou schéma fourni #
- Format (informatique) : txt, .csv, XML; HTML
- Personnalisable, associé à des choix de contenu

Edition graphique (terminogramme, schémas fléchés)

Gestion de l'activité de création et maintenance

- Noms des gestionnaires, gestion des accès
- Gestion des dates : création, modifications
- Outils d'aide aux tâches : marquage de lots de descripteurs, etc.

Statistiques sur le thésaurus

- Nombre de domaines, microthésaurus, termes (par type) #
- Nombre total de termes
- Nombre de termes par catégories (domaines, microthésaurus)
- Nombre de notes, par types de note
- Nombre de termes par niveaux hiérarchiques
- Nombre de descripteurs ayant une relation avec un autre terme, pour chacun des types de relation (hiérarchique, associative, équivalence)

Statistique sur son utilisation

- Occurrence des termes
- Co-occurrence des termes

Ergonomie

- Facilité d'installation #
- Langue de l'interface du logiciel #
- Interface utilisateurs en accord avec les règles actuelles de l'art
- Indicateur de position (fil d'Ariane)
- Glisser/déposer : terme, branches, relations #
- Multifenêtrage possible (visibilité simultanée de plusieurs zones du graphe)
- Couverture et qualité de la documentation, de l'aide en ligne #

Utilisation : recherche et indexation, consultation du thésaurus

Dépend fortement de l'applicatif développé. Voir aussi Ergonomie

- Prise en compte implicite des équivalents
- Déplacement dans l'index des termes :
- index permuté avec occurrences
- avec présentation des synonymes et occurrences
- Navigation par arborescence au sein du réseau
- Sélection de plusieurs termes
- Fonction « panier » pour collecter les termes
- Lien direct entre les termes et les notices consultables
- Pas de contrainte de casse (lettre capitale et minuscule)
- Pas de contrainte d'accent
- Accepte les troncatures ou mots partiels
- Indicateur de position (fil d'Ariane)
- Autopostage, ascendant ou descendant, niveau d'extension paramétrable par l'utilisateur
- Autopostage automatique, géré par l'administrateur
- Possibilité d'associer un thésaurus indépendant des données documentaires

Critères techniques du module ou logiciel

Critères communs à tout logiciel de gestion de données / bases de données

- Mono-, multi-utilisateur; client/serveur et Internet
- Caractères supportés
- Plateforme de consultation, de gestion
- Base(s) de données supportées #

Critères proposés par L. Will [7]

documents du fonds déjà indexés par ces termes, soit les notes d'application. Mais il n'a pas fini : il lui faut poursuivre sa navigation pour une autre notion (ou pour les autres notions) qu'il souhaite combiner pour forger l'indexat[16]. Une fonction de type « panier » ou « sélection multiple » est alors nécessaire [*figure 1*].

Des contraintes ergonomiques, comme celles liées à la fenêtre-écran qui réduit la vision globale du réseau sémantique, freinent considérablement les plus fervents à la tâche. Si l'indexeur connaît déjà les descripteurs, le module contrôle automatiquement la présence du terme saisi par rapport au thésaurus. Dans le cas où le terme sélectionné est un non-descripteur, celui-ci est automatiquement remplacé par le descripteur. Le module autorise également l'autopostage descendant, c'est-à-dire l'insertion automatique des termes spécifiques du descripteur sélectionné[17], ou ascendant.

Lorsque la notion recherchée n'appartient pas au référentiel terminologique, deux cas de figure peuvent se présenter. Dans le premier, la politique d'indexation a pu définir pour cette notion le renvoi vers deux descripteurs. Ce renvoi de coordination (« USE… », « AND… »), prévu par les normes et visible dans les thésaurus les plus anciens, n'est que rarement implémenté dans les logiciels et de fait totalement omis aujourd'hui lors de la conception des thésaurus, voire dans les manuels d'indexation. L'autre cas suppose l'ajout par l'indexeur d'un candidat descripteur dans une zone ou avec un balisage *ad hoc*. L'indexeur peut également souhaiter établir une annotation complémentaire sur l'usage d'un descripteur.

Les difficultés tiennent ici plus à une ergonomie contraignante, et cette rigueur imposée sera surtout suivie et acceptée par des indexeurs professionnels. Mais que proposer à l'indexeur occasionnel, professionnel de l'I-D aux multiples tâches mais également auteur de ressources à qui il est aujourd'hui demandé de participer à ce travail de référencement « à la source » ?

▼ Langage d'expansion à la recherche

Si les choses semblent claires en ce qui concerne l'utilisation à l'indexation d'un thésaurus ou plus largement de vocabulaires contrôlés, l'usage du thésaurus documentaire à la recherche est franchement controversé. Malgré des progrès indéniables depuis une décennie, l'ergonomie[18] des activités liées à la recherche dans des ressources indexées à l'aide d'un thésaurus reste encore très traditionnelle, avec une prise en compte très limitée des pratiques et des difficultés réelles des utilisateurs.

L'étude des applications qui ajoutent ici une liste de synonymes, là un thésaurus de langue ou autres référentiels terminologiques, semble nous montrer qu'est bien remise en cause l'utilisation « jumelle » d'un même thésaurus normatif pour la recherche et pour l'indexation.

Pour la recherche, nous pouvons distinguer deux situations extrêmes.

Le thésaurus doit être consultable pour certains utilisateurs lors de recherches ponctuelles précises, mais aussi pour la construction de profils. Ces profils sont élaborés par des documentalistes dans le cadre d'une offre globale[19] ou d'une prestation personnalisée, mais également par des utilisateurs, classiquement des chercheurs mais aussi des juristes ou des chargés de produits pour leur veille, qui souhaitent conserver une certaine autonomie pour la gestion de leur programme de recherche d'information. Peut-être souhaitent-ils aussi conserver une technique de recherche d'information avec laquelle ils sont familiers et qu'ils ont fini par bien maîtriser ?

Dans ce contexte, les documentalistes et surtout les utilisateurs finals doivent pouvoir explorer le thésaurus avec leurs propres mots sans s'y perdre. Pour répondre à ces pratiques, les points d'entrée dans les thésaurus ont été multipliés, augmentant ainsi le nombre de non-descripteurs. Aux fonctions de consultation du thésaurus proposées à l'indexeur [*voir ci-dessus*], s'ajoute un outillage dont les spécifications relèvent plus de l'ergonomie de l'interface, comme le « fil d'Ariane » ou la présentation des index permutés de tous les termes. Exception faite des logiciels développés dans certains environnements comme les bibliothèques mais aussi parfois dans l'audiovisuel, les thésaurus sont souvent consultables à la recherche et associés à la construction de profils personnalisés. Mais sans aucune aide pour l'interrogateur ni même le plus souvent sans panier pour sélectionner plusieurs descripteurs, ces fonctions se présentent quasiment sous la forme proposée sur les postes client/serveur des années quatre-vingt.

Mais la plupart des utilisateurs souhaitent passer outre cette phase complexe de formulation d'une requête à partir de la sélection de termes dans des vocabulaires contrôlés[20], préférant porter leur attention sur la fouille du lot résultat. Ceci est d'autant plus vrai lorsque le portail propose des accès à de nombreuses ressources dans lesquelles les termes à utiliser pour désigner une même notion peuvent être différents de l'une à l'autre. Ce contexte n'est pas nouveau puisqu'il correspond au fondement même des serveurs de bases d'information professionnelle, le web invisible.

Quel rôle attribuer au thésaurus dans ce contexte ? Une solution consiste à ne pas l'utiliser dans sa forme relationnelle et à n'exploiter que les *formes équivalentes*. On n'exploite ici, dans une recherche dite *plein texte*, que les descripteurs de la notice auxquels sont ajoutés les non-descripteurs associés à chacun des descripteurs. Dans ce schéma, les indexats des notices constituent une

zone de recherche par mots, zone complémentaire au titre et au résumé lorsque ce dernier existe. Les logiciels documentaires traditionnels assurent ce type de recherche, d'autres outils de gestion électronique de documents et de gestion de contenu web (souvent appelés un peu abusivement thésaurus) offrent également la possibilité d'intégrer des listes de synonymes pour améliorer la recherche, en l'occurrence le rappel.

Mais, si elle a le mérite de la simplicité, cette solution ne résout pas tous les problèmes rencontrés au cours d'une recherche d'information, surtout dans le cas d'une recherche dans une base de références documentaires.

En particulier, le problème du silence induit par le décalage de niveau hiérarchique entre les notions présentes dans la question et celles représentées dans l'indexat des documents (ou le titre) reste entier. Ainsi un document traitant de « résultats d'études épidémiologiques sur le cancer en France, Espagne et Italie », indexé par ces trois noms de pays, ne sera pas proposé en réponse à une question sur « les problèmes d'épidémiologie du cancer dans l'Union européenne ». La pratique professionnelle d'indexation aux notions précises, sans surindexation au générique (ici Union européenne) pour ne pas générer du bruit, ne prend son sens que dans le cadre de l'utilisation de fonctions d'expansion automatique, ici ascendante lors de la recherche.

Or ces fonctions d'expansion automatique à partir des thésaurus sont peu mises en œuvre[21] ou alors uniquement dans certains applicatifs dédiés [10]. La présence de ressources multiples traitées avec des vocabulaires contrôlés différents rend ici caduque l'utilisation d'un langage d'indexation (lequel choisir ?), mais donne du poids aux tenants d'un métalangage construit spécifiquement pour l'accès à l'information. Ces fonctions d'expansion automatique sont par contre constitutives des logiciels en langage naturel et participent de leur succès.

16 Indexat : ensemble des termes ou indices issus de l'indexation d'un document et associé à ce document.
17 Nous pouvons citer le cas classique d'indexation au nom des différents pays de l'Union Européenne, à partir d'un traitement par la machine sur le descripteur « Union européenne ».
18 Nous faisons ici référence à l'ergonomie fonctionnelle et logicielle, et non à l'ergonomie dite de « surface ».
19 Les utilisateurs s'abonnent à un profil et reçoivent les résultats (diffusion sélective de l'information ou DSI) *via* un flux RSS, par courriel ou directement sur leur espace personnel sur le portail.
20 Sur les difficultés d'utilisation des thésaurus, voir : Annette Béguin, « Usages du thésaurus et développement de la pensée », *InterCDI*, mars-avril 1999, n° 158, ou : Élisabeth Nasse-Kolmayer, *Contribution à l'analyse des processus cognitifs mis en jeu dans l'interrogation d'une base de données documentaires*, thèse, 1997.
21 Les difficultés de mise en œuvre de ces techniques automatiques d'expansion (autopostage) viennent aussi d'une construction trop lâche des relations hiérarchiques au sein du thésaurus, qui rendent l'autopostage de qualité trop variable.

Si le modèle de la formulation sans contrainte de la question (mais ici sans traitement linguistique) se rapproche des modèles proposés par les moteurs de recherche, l'ordonnancement des résultats et leur présentation sont restés classiques. Le thésaurus ou au moins les indexats pourraient être utilisés en aval pour pondérer les résultats de la recherche à partir des occurrences ou co-occurrences des descripteurs, ou plus simplement pour organiser ce lot résultat par rapport aux thèmes ou domaines d'appartenance des descripteurs. Ces solutions largement mises en œuvre par les outils automatiques, linguistiques ou statistiques relèvent pour le moment d'applicatifs spécialisés. Dans cette même catégorie d'accès à l'information *via* un métalangage, citons les produits beaucoup plus récents qui s'appuient sur des schémas de représentation de la connaissance, plus structurés et formalisés que les thésaurus, comme les *topics maps* ou les ontologies.

▼ Représentation graphique

Si de nombreux professionnels de l'information-documentation n'hésitent plus à réviser parfois en profondeur les interfaces de recherche en vue de simplifier la tâche de l'utilisateur, cette remise en cause n'atteint pas encore les modalités d'interaction proposées qui restent essentiellement textuelles. La représentation graphique ou visuelle de l'information utilisée dans les années soixante sur papier pour les schémas fléchés des thésaurus, puis sur ordinateur dans les activités de veille[22], a pris un nouvel essor sur le Web pour la recherche et la présentation des résultats, avec par exemple le moteur Kartoo dès 2001. Différentes applications sont aujourd'hui visibles *via* l'internet : l'interface DeweyBrowser, développée dans le cadre d'un projet de recherche par l'On Line Computer Library Center (OCLC), s'appuie sur une présentation originale de la classification de Dewey[23] ; le logiciel StarTree d'Inxight est utilisé par la National Science Digital Library (NDSL) pour naviguer au sein de ses collections par sujets[24] ; la Queens Library de New York offre une représentation graphique des sujets du lot de documents résultat d'une recherche avec le produit Aquabroswer[25]. On peut également citer le projet d'Ebsco[26] avec le produit Grokker, l'application VisualThesaurus[27] ou le système ►

22 Les professionnels de l'I-D qui travaillent dans ce type de dispositifs connaissent depuis de nombreuses années déjà les outils cartographiques associés à la fouille de texte.
23 http://deweyresearch.oclc.org/ddcbrowser/wcat
24 http://nsdl.org/browse/ataglance/browseBySubject_netmac.html (rubrique Browse at a Glance).
25 http://aqua.queenslibrary.org/
26 Ebsco avec le moteur Grokker (http://support.epnet.com/uploads/CustSupport/Images/ReleaseInfo/VisualSearch.gif).
27 Visualisation graphique du thésaurus de langue (www.visualthesaurus.com/).

d'interrogation et d'exploitation du catalogue à la bibliothèque universitaire de Paris-8[28].

Ressource terminologique et sémantique autonome

Si la situation est controversée en recherche documentaire, les applications de consultation de thésaurus en tant que ressource terminologique autonome se multiplient. Les fonctionnalités proposées et l'ergonomie mise en œuvre par ces interfaces font en France l'objet de développements spécifiques. Elles proposent les modes d'accès courants au thésaurus avec des ergonomies variables : requête par mot libre, navigation et sélection au sein d'un index alphabétique, parfois d'un index permuté (BDSP, MOTBIS), et navigation par arborescence.

Deux orientations se dessinent, suivant deux axes déjà évoqués :
- une orientation « référentiel normatif et terminologique » avec des sites dédiés comme ceux de la FAO[29] ou de la NLM[30] ;
- une orientation « aide à la recherche », avec des interfaces autorisant soit une connexion directe à la base documentaire qui est à l'origine du thésaurus (Eric ou BELIT), soit le lancement de la recherche sur des moteurs avec une logique de langage de recherche. L'application proposée par le *Thésaurus canadien d'alphabétisation* permet à l'utilisateur de naviguer dans le thésaurus bilingue français/anglais, de sélectionner un seul terme, puis de lancer automatiquement une recherche au choix sur Yahoo! ou sur Google. L'éditeur du logiciel spécialisé MyThesaurus[31] a construit une interface plus complexe permettant de sélectionner plusieurs termes et d'élaborer une requête, avant de lancer celle-ci sur plusieurs moteurs de recherche.

Un vocabulaire à concevoir et à maintenir

Les orientations professionnelles précédemment évoquées nous conduisent à réviser le terme usuel de *gestion de thésaurus*, et en particulier à distinguer :
- les activités de maintenance au fil de l'eau de celles de conception ou de reconception si une mise à jour importante est prévue ;
- les situations où le vocabulaire est articulé à une base documentaire de celles où le développement du (ou des) vocabulaire(s) contrôlé(s) est autonome par rapport à son (leur) exploitation (indexation, recherche) dans différents dispositifs.

▼ Maintenance du thésaurus et index des bases documentaires

Dans le schéma documentaire traditionnel (le même outil servant à l'indexation et à la recherche), les modifications apportées au thésaurus peuvent avoir de fortes répercussions sur les indexats des notices. C'est une phase délicate que l'on aimerait réaliser, si ce n'est automatiquement, au moins avec un appui plus fort de l'informatique. Précisons d'emblée pour rassurer un grand nombre de professionnels : dans les systèmes documentaires actuels, les modifications de relations associatives et hiérarchiques sont sans incidence sur les index des bases documentaires. Les relations appartiennent au thésaurus et, dans ce modèle documentaire traditionnel, elles ne sont pas transférées dans les notices.

Les problèmes surgissent lors de modifications apportées aux termes et aux relations d'équivalence. Une modification de la forme rédactionnelle d'un terme ne devrait théoriquement pas poser de problème majeur et conduire automatiquement à une mise à jour des indexats de l'ensemble des notices concernées. Mais, lorsqu'un descripteur est supprimé ou que son statut est modifié (de descripteur à non-descripteur, et vice versa), tous les logiciels n'offrent pas les mêmes solutions, certains d'ailleurs n'en proposant aucune. D'autres encore suppriment bien le descripteur du thésaurus, mais ne gardent aucune trace de son existence passée.

▼ Gestion de candidats descripteurs

Dans le contexte de la promotion faite actuellement à une certaine forme d'indexation par l'utilisateur (folksonomie ou indexation collaborative), il est intéressant d'étudier les pratiques des professionnels et les fonctionnalités logicielles existantes. Les traditionnels « mots clés des auteurs », dans la documentation scientifique et technique, sont selon le cas : intégrés à la notice comme un complément du titre ou du résumé, ou bien immédiatement pris en charge par les documentalistes dans le thésaurus ou une liste des candidats, ou encore… totalement occultés.

Il convient ici de distinguer d'une part la présence d'un champ autonome (« élément de données », dirions-nous aujourd'hui) que l'on pourrait envisager ouvert aux utilisateurs/lecteurs, et d'autre part le lien de cet élément de données avec une liste de valeurs ouverte, indépendante du thésaurus ou gérée en son sein. Cette architecture est fonction des possibilités offertes par le logiciel ou par l'application d'associer un même thésaurus ou une même partie de thésaurus à plusieurs champs d'une même base. Dans le cas où la liste est gérée au sein du thésaurus, ces candidats descripteurs peuvent être rapidement traités aussi bien sur le plan de leur statut au sein du thésaurus (nouveau descripteur ou équivalent) qu'au niveau du traitement de la notice.

▼ (Re)conception d'un thésaurus existant

La production documentaire (*i.e.* la production de références documentaires) en réseau et la disponibilité toujours plus grande de sources et ressources documentaires reposent la question du territoire de la base documentaire produite en interne.

Ces projets de réingéniering documentaire sont à l'origine d'un travail en profondeur sur les vocabulaires contrôlés. Ces refontes de langages, la création de langages dédiés à la recherche ou les travaux sur la concordance entre langages d'indexation supposent des fonctionnalités particulières pour lesquelles les modules des progiciels documentaires peuvent s'avérer insuffisants. Une première condition est de pouvoir travailler sur un thésaurus déconnecté d'une base documentaire. D'autre part ces activités réclament des fonctions de personnalisation, une grande souplesse quant au modèle de données avec la création de type de relations ou de catégories de termes particuliers[32], ou encore la présence d'un outillage pour organiser et conduire les travaux entre gestionnaires et administrateurs, comme le marquage de termes à la volée.

▼ Contrôles

La spécificité des modules thésaurus tient pour une grande part dans les contrôles automatiques des relations sémantiques. Il n'en reste pas moins que des problèmes (contrôles inopérants ou partiels, absence d'information sur la raison des rejets) sont régulièrement mentionnés par les professionnels, les obligeant à un audit périodique en parallèle. De plus les développements récents autour des technologies du Web mettent sur le marché des applications intéressantes, mais dont la principale faiblesse porte justement sur ces contrôles.

▼ Import

La nécessité d'importer un thésaurus signe des changements importants : il s'agit de récupérer l'intégralité d'un vocabulaire contrôlé, les données actives mais aussi l'historique ou les correspondances avec d'autres vocabulaires (comme l'appartenance d'une liste annexe d'un thésaurus à une nomenclature métier de l'entreprise) et, dans le cas d'un changement de logiciel, les données documentaires. Plusieurs problèmes de nature différente surgissent alors.

Le modèle du thésaurus ou certaines caractéristiques sont différentes : le thésaurus est polyhiérarchique et le logiciel que l'on vient d'acquérir ne prend pas en compte cette caractéristique ; ou encore la notion de terme préférentiel n'existe pas dans le thésaurus, tous les termes étant *a priori* équivalents, mais le nouveau logiciel l'exige ; enfin la longueur permise pour les termes dans le nouveau logiciel est faible. Un travail d'importance variable reste ici à réaliser avant le transfert, avec en toile de fond le report de ces modifications sur les indexats des notices documentaires. Si la situation s'arrange pour la polyhiérarchie et le terme préférentiel, les différences entre logiciels dans la gestion du multilinguisme, la longueur autorisée pour les termes ou dans le mode d'assignation des termes à des catégories perdurent. Les fichiers traces des données en erreur, produits par le logiciel lors de l'import, sont d'un grand secours et facilitent la requalification des vocabulaires.

La synchronisation avec les données documentaires peut également constituer un point critique, si le logiciel ne gère pas la notion de thésaurus actif et non actif ou hors ligne.

Les normes sur les thésaurus ainsi que les standards plus récents comme SKOS[33] ou zThes n'intègrent pas toutes les données nécessaires à un administrateur de vocabulaires. Il faut donc toujours prévoir une part de développement ou *a minima* de paramétrage.

▼ Personnalisation

La personnalisation est une fonction doublement intéressante : elle permet bien sûr de créer un modèle de thésaurus adapté aux besoins de l'utilisateur dans le cadre fourni par les normes. Par exemple, une relation d'instance pourrait être pertinente pour les listes de noms d'organismes, relation à laquelle l'autopostage automatique ne serait pas appliqué. Mais ces fonctions de personnalisation offrent également une grande souplesse et un appui aux travaux – de nature terminologique – de reconception des thésaurus ou à des travaux plus étendus au sein des organismes.

▼ Multilinguisme

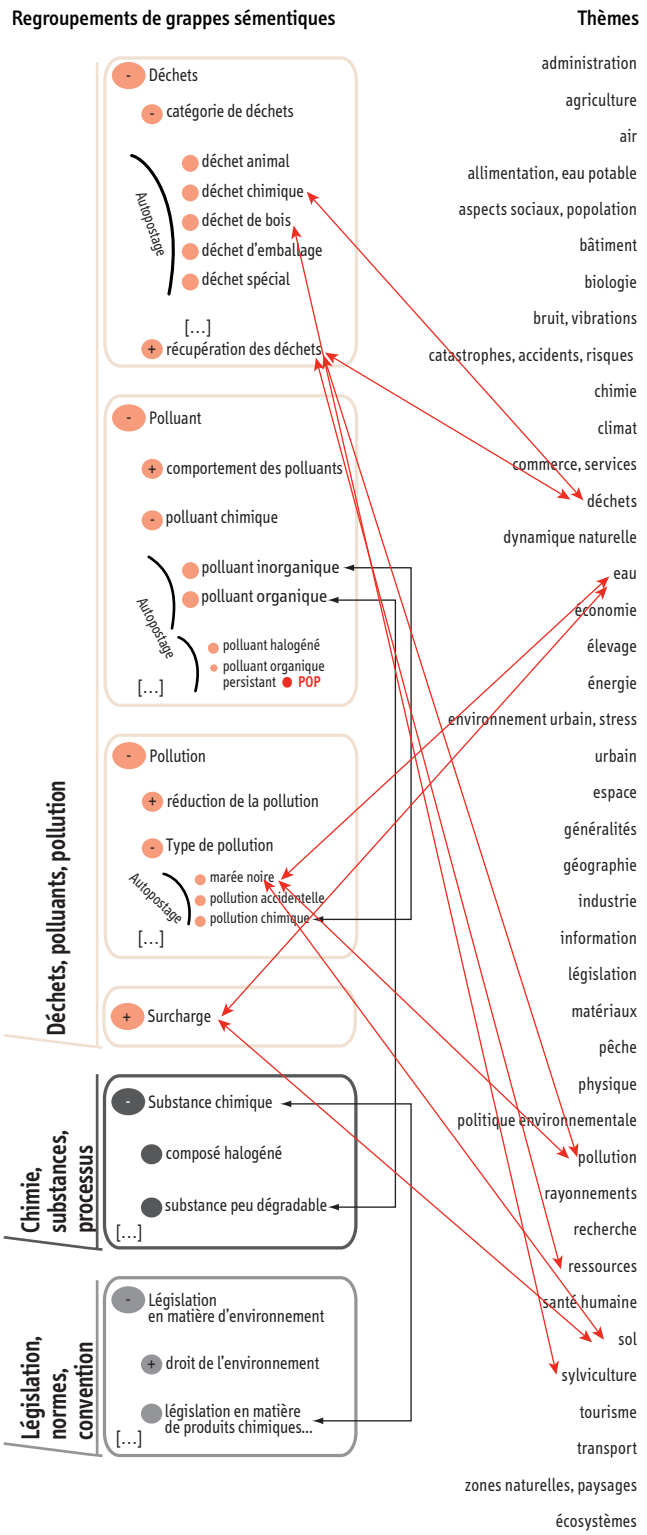
Que ce soit dans le cadre d'entreprises ou d'organismes à couverture internationale (intranet *a minima* bilingue), de pays plurilingues ou d'organismes voulant toucher des populations d'origines linguistiques variées, cette question revêt une importance et des formes variées.

Sur le plan des thésaurus, le multilinguisme est vu sous l'angle de la relation d'équivalence, ici linguistique, à un descripteur du thésaurus dans la langue source. En plus de la question de l'encodage informatique avec une norme ISO 10646 (ou Unicode), devenu incontournable, il convient de prêter attention aux modèles proposés par les logiciels pour traiter cette question, certains d'entre eux pouvant conduire à des difficultés de mise en œuvre. Si l'on souhaite donner des accès aux utilisateurs dans chacune des langues traitées, il faut pouvoir intégrer des synonymes pour chacune des versions linguistiques et pas uniquement pour la langue source.

Dans certains environnements multilingues, on peut aussi vouloir proposer des catégories ►

28 Qui est plus qu'une simple représentation graphique (http://visualcatalog.univ-paris8.fr/vc2/).
29 www.fao.org/aims/ag_intro.htm
30 www.nlm.nih.gov/research/umls/
31 www.manggix.com/mythesol/index.php
32 Voir OTAREN pages 89-92.
33 Voir SKOS page 75.

FIGURE 2 – RESEAU SEMANTIQUE D'UN THESAURUS



adaptées à chaque environnement culturel et linguistique. Citons le cas classique de la place des thèmes « Sport » et « Loisirs » classés dans le domaine « Culture » dans des organisations d'origine anglo-saxonne (Unesco), alors que la logique française classe bien « Sport » dans « Loisirs », mais ce dernier dans le domaine « Mœurs », avec une simple relation associative à « Culture populaire », lui-même classé dans « Culture » ; mais, tant en contexte anglo-saxon que français, « Culture » est pourtant associé à « Vie culturelle ».

▼ Modèle de données et normes

Pour concevoir les modules « thésaurus » et leur articulation avec les bases de données pour l'indexation et pour la recherche, les éditeurs de logiciels sont partis du modèle de données proposé par les normes ISO de 1986, ou plus fréquemment en France de la norme Afnor de 1981. Force est de constater que les résultats en termes de contraintes et de fonctionnalités sont assez variables. Ces variations ont plusieurs origines.

D'une part, les protagonistes (éditeurs mais aussi documentalistes) peuvent avoir une compréhension différente de certains aspects de ces normes, en particulier la place des domaines ou des champs sémantiques par rapport aux descripteurs[34] [*figure 2*] et l'intérêt ou non de la spécialisation des relations hiérarchiques[35] ou du statut des non-descripteurs. Les possibilités offertes par les normes peuvent aussi ne pas être toutes prises en charge, certaines d'entre elles étant explicitement non imposées dans les normes, d'autres étant jugées trop complexes (comme le renvoi d'un non-descripteur vers deux descripteurs à coordonner) et n'étant pas demandées explicitement par les utilisateurs.

D'autre part, le degré d'importance accordé ou non à ces outils linguistiques par rapport aux autres fonctions du logiciel (portail, RSS, etc.) limite également le temps et les moyens consacrés à ces modules. Enfin, les normes de thésaurus ont adopté des styles de présentation moins formels que ceux choisis pour la présentation d'autres normes ou de schémas/DTD, plus complexes aussi à mettre en œuvre. En particulier, les normes et donc les logiciels qui en découlent n'intègrent pas les éléments de données nécessaires à l'utilisation[36] ou à la maintenance des vocabulaires : dates de création et de mises à jour, noms de gestionnaires, source ou origine d'un terme, lien avec d'autres langages.

Certains éditeurs ont accepté de sortir du cadre simplificateur de la norme stricte pour répondre à des besoins particuliers, modifiant ainsi le modèle proposé par la norme. C'est le cas pour la relation d'équivalence historique et les travaux de concordances entre langages.

Certains logiciels proposent une *relation d'équi-*

valence historique pour des termes obsolètes. Cette relation gère ainsi le lien entre deux termes, Areva / Cogema, Framatome ou France Telecom / Orange. Alors que la pratique faisait basculer le plus ancien terme en équivalence du plus récent avec suppression de l'ancien dans les notices[37], ces deux termes peuvent coexister dans le thésaurus mais également dans la base : les documents plus anciens indexés avec « CNPF » doivent conserver cet indexat, le terme garde donc un statut de descripteur. Cette fonction est valable pour tous les changements de noms propres (entités nommées) d'organismes, de noms de produits, etc., et peut être utilisée largement pour d'autres vocabulaires contrôlés. Ces descripteurs obsolètes ne peuvent plus être utilisés à l'indexation, mais ils apparaissent dans les différentes listes à la recherche ou à l'édition.

La *concordance avec des descripteurs d'un autre thésaurus* concerne plus particulièrement les thésaurus de recherche, comme dans le cas d'OTAREN. On peut aussi citer le *Thésaurus d'éthique des sciences de la vie* [*figure 1*], qui établit des concordances avec quatre thésaurus : le *Bioethics Thesaurus* du Kennedy Institute of Ethics (lettre B), l'*Euroethics Thesaurus* (lettre E) dont le développement a été suspendu, le *MeSH* de la NLM (lettre M avec le code du descripteur) et le *Thésaurus d'éthique des sciences de la vie et de la santé* (lettre I), développé par le Centre de documentation en éthique des sciences de la vie et de la santé.

Les organismes souhaitant gérer tout à la fois les thésaurus et des terminologies suivant les règles normalisées pour ces deux catégories d'outils linguistiques devront de toute évidence réfléchir au recouvrement entre ces différents modèles[38].

34 L'architecture des données peut privilégier une collection de termes hiérarchisés par grappes, relativement autonomes vis-à-vis des catégories de classement (un thésaurus peut d'ailleurs ne pas être composé ni de domaine ni de champ sémantique, ou au contraire favoriser le classement des termes par domaine, les points d'entrée étant alors les catégories).
35 La norme ISO de 1986 codifie par exemple la relation d'instance, ce que ne fait pas la norme française.
36 Le profil zThes, par exemple, intègre des éléments d'historisation des descripteurs.
37 Ou bien la conservation des deux termes reliés par une relation associative, avec le risque que l'un ou l'autre de ces deux termes ne soit pas connu ni donc utilisé à la recherche.
38 La norme ISO 16642 - TMF (Terminological Mark-up Framework) fournit un cadre pour la représentation des bases de données terminologiques en XML (Termsciences : www.termsciences.fr/rubrique.php3?id_rubrique=24). Un exemple : www.termisti.refer.org/dessisti2005.dtd
39 Nous ne traitons pas ici des solutions proposées par les serveurs de banques de données, celles-ci ne pouvant être acquises dans le cadre d'un projet d'entreprise.

2 Les familles d'outils logiciels

Nous proposons dans cette section un rapide panorama des familles de logiciels centrés sur les thésaurus[39].

Modules « thésaurus » de logiciels de gestion et recherche documentaires

Le module spécialisé associé à un logiciel de gestion et recherche documentaires est la situation la plus fréquemment rencontrée sur le terrain en France depuis le début des années quatre-vingt[40]. Ce module permet d'exploiter à la recherche et à l'indexation un ou plusieurs thésaurus ou d'autres types de vocabulaires moins complexes, et de réaliser les tâches de maintenance au fil de l'eau [3].

Les caractéristiques associées aux trois activités de gestion, d'indexation et de recherche, présentées dans le tableau A pages 46-47, sont amplement couvertes par la plupart des LGRD, d'Alexandrie à Cindoc ou JLB-Doc en passant par Ever, DIP, Cadic, avec des ergonomies variables qui ne sont pas sans impact sur l'efficacité de la tâche. Toutefois, certaines différences fonctionnelles et qualitatives entre les modules pourraient mener à certaines impasses pour des projets plus ouverts (interopérabilité entre langages, recherche multisources, etc.).

Attachés au modèle traditionnel « indexation – thésaurus – recherche », ces produits se sont orientés vers des partenariats avec des éditeurs spécialisés, d'abord pour l'indexation automatique des ressources numériques, puis plus récemment avec des logiciels linguistiques pour répondre aux nouvelles exigences des utilisateurs pour la recherche documentaire. Aujourd'hui, les offres sont structurées autour d'un ensemble de briques utiles au développement d'interfaces de recherche performantes : langages contrôlés, indexation automatique statistique et recherche en langage naturel. Dans ce contexte, la base documentaire et le thésaurus sont plutôt destinés à la production documentaire.

Interfaces autonomes d'exploitation d'un thésaurus

Le thésaurus en tant que référentiel terminologique est consultable à travers des interfaces dédiées issues de développements spécifiques ►

40 Parmi les évolutions récentes de ces produits, on peut citer une certaine normalisation informatique, d'abord par le passage aux technologies des SGBR relationnels et ODBC, puis plus récemment par la prise en compte des technologies des services web. Cette normalisation apporte une plus grande autonomie au module Thésaurus par rapport à la base de données d'origine. Toutefois ces modules ne sont pas totalement indépendants de l'architecture informatique de la base documentaire sur laquelle ils reposent (ils ne sont d'ailleurs pas commercialisés de façon autonome), même lorsque le logiciel est proposé sous un autre SGBDs comme Oracle.

Tableau B - LOGICIELS DE CREATION ET MAINTENANCE DE VOCABULAIRES CONTROLES

Nom du produit et nom de la société (si différent)	Pays éditeur	Localisation Internet
a.k.a. de Synercon Management Consulting	Australie	http://a-k-a.com.au/aka_classification/index.htm
Amicus® Thesaurus tool	Canada	www.amicuscom.com/
Cognatrix de LGOSystem	Australie	www.lgosys.com/products/Cognatrix/index.html
domainREUSER	Espagne	www.reusecompany.com/producto.aspx?id=13
IC INDEX 5.0	Allemagne	www.agi-imc.de/
MIDOSThesaurus de Progris	Allemagne	www.progris.de/
MultiTes de MutiSystems	États-Unis	www.multites.com/
MyThesaurus de Manggix	France	www.mythesaurus.fr/
Stride de Questans	Royaune-Uni	www.questans.co.uk/p100l2.html
Synaptica KMS de Factiva	États-Unis	www.factiva.com/products/taxonomy/ synaptica.asp?node=menuElem1511
TemaTres de R020 *libre, gratuit*	Argentine	www.r020.com.ar/tematres/index.en.html
Term Tree, ACS	Australie	www.termtree.com.au/html/termtree.html
The32W de l'University of Western Ontario *libre, gratuit*	Canada	http://publish.uwo.ca/%7Ecraven/freeware.htm
Thesaurus Master de DataHarmony	États-Unis	www.dataharmony.com/products/tm.htm
THESmain	Autriche	www.cedar.at/wgr_home/
Trias Politica Thesaurus Builder	Iran, Pays-Bas	www.thesaurusbuilder.com/index.asp
WebChoir	États-Unis	www.webchoir.com

(BDSP, MOTBIS, GEMET). Les différences portent plus sur l'ergonomie que sur les fonctionnalités. Dans les pays anglo-saxons, les logiciels de conception et maintenance de thésaurus sont plus fréquemment utilisés, offrant des caractéristiques et une ergonomie moins variables[41]. Nous pouvons également citer dans cette catégorie le logiciel autonome MyThesaurus Online qui permet d'interroger des ressources sur le Web, dont les moteurs de recherche, à partir de la consultation d'un thésaurus.

Logiciels de conception et de maintenance de thésaurus

Nous le disions en introduction, cette famille de logiciels[42] [*voir ci-dessus*], pourtant de la première heure[43], reste peu connue en France, les pratiques favorisant en cas de besoin des applications « maison »[44]. La situation dans les organisations change beaucoup, et le rachat en 2005 par Factiva d'un des plus importants logiciels spécialisés américains, Synapse de la société Synaptica[45], fournit un indice de l'importance accordée aux vocabulaires contrôlés et à leur gestion autonome dans la stratégie d'un serveur d'information professionnelle.

Ainsi le développement de référentiels terminologiques dans les organismes peut nous conduire à nous intéresser aux logiciels indépendants. Centrés sur les corpus terminologiques, ces progiciels sont dotés d'une palette de fonctionnalités de (re)conception et de suivi mieux adaptées à ces nouvelles situations : richesse des fonctions d'import, souplesse dans les modèles de données et les possibilités de personnalisation, insertions ponctuelles (termes, termes hiérarchisés, notes) par simple couper/coller, marquages personnalisés, gestion des tâches, statistiques personnalisables, etc.

Perspectives

Si les logiciels (modules ou progiciels) permettant à la fois la gestion et l'exploitation de thésaurus sont anciens et nous renvoient à une image traditionnelle de l'accès à l'information, les applications, expériences ou projets en développement depuis quelques années montrent une situation beaucoup plus variée, avec des orientations fortes sur trois plans.

• Une distinction nette entre système d'indexation et système d'interrogation, avec en toile de fond une réévaluation de la place des bases de références documentaires par rapport aux fonds de documents numériques. Dans ce cadre, de nouveaux outils apparaissent, comme ITM de Mondeca capable d'intégrer des thésaurus existants « *dans un système unifié de représentation des connaissances, utilisant un modèle générique de type Topic Map*[46] ». L'implication des professionnels de l'information-documentation dans ces familles d'outils ne fait aucun doute mais impose une révision de certaines de nos pratiques [9].

Langue de l'interface
anglais
anglais
anglais
espagnol
allemand
allemand
anglais
français
anglais
anglais
espagnol, portugais, anglais
anglais
anglais
anglais
allemand
anglais
anglais

• Un renforcement du suivi et de la maintenance des vocabulaires contrôlés, avec un usage plus large et plus autonome de logiciels spécialisés.

• Un développement d'activités liées à la conception de vocabulaires (pour tel ou tel dispositif documentaire, dans le cadre de la participation à tel réseau de dépouillement, pour l'établissement d'une terminologie au sein de l'entreprise, etc.) ou à la conception d'interfaces d'accès à l'information.

Dans tous les cas, pour pouvoir concevoir des langages opérationnels et interopérables ou imaginer des interfaces adaptées, il semble indispensable de mettre en place des programmes de suivi et d'évaluation de systèmes pour recueillir le matériau nécessaire à ces développements.

41 Thésaurus de la Banque mondiale avec Multites (www.multites.com/wb/).
42 Terminologie anglo-saxonne : *thesaurus management software*.
43 MultiTes existe depuis 1982.
44 Le logiciel GTI utilisé pour la construction d'OTAREN a été réalisé par le CNDP de Poitiers. Voir pages 89-92.
45 Factiva Synaptica Knowledge Management System, sur le site de Factiva (www. factiva.com).
46 www.mondeca.com/fr/referentiel.htm

Ressources bibliographiques

Caractéristiques de modules « thésaurus », d'interfaces ou de logiciels indépendants

[1] BRIOT, Bernadette, Knapen, Étienne. « Test de gestionnaires de thésaurus pour la terminologie ». *La banque des mots*, 1999, n° 58, p. 31-50 [Présentation du fonctionnement puis grille de comparaison appliquée à cinq modules logiciels dont quatre ne sont plus disponibles]

[2]]ACSO, Peter. « Using Controlled Vocabulary ». *In* : « "Savvy searching" columns », paru dans *Online Information Review*
« Content part ». 2003, vol. 27, n° 4, p. 284-286
« Part II - Software Issues ». 2003, vol. 27, n° 5, p. 359-363
« Part III - Query Mapping and Thesaurus Term Suggestion ». 2003, vol. 27, n° 6, p. 446-450
[Trois articles courts et critiques sur les interfaces d'interrogation des bases de données en ligne]

[3] LACHANA, Évanghélia. Mise à jour d'un thésaurus : éléments et propositions de méthode à partir de la mise à jour du thésaurus du centre de documentation sur la formation et le travail du Cnam. Mémoire INTD, octobre 2001 NB pages [Sur la méthode et l'organisation du travail de reconception du thésaurus et de mise à jour de la base sous un progiciel documentaire]

[4] MIDDLETON, Michael. Controlled vocabularies. 2007. http://sky.fit.qut.edu.au/~middletm/cont_voc.html [Site aux ressources variées : listes de vocabulaires, de logiciels, bibliographie]

[5] RIESLAND, Melissa A. « Tools of the trade: vocabulary management software ». *Cataloging & Classification Quarterly*, 2004, vol. 37, n° 3-4, p. 155-176 (DOI : 10.1300/J104v37n03_10)

[6] ROHO, Cécile. « La gestion automatisée des thésaurus : étude comparative de logiciels ». *Documentaliste - Sciences de l'information*, 1987, vol. 24, n° 3, p. 103-108 [Présentation de dix modules logiciels, dont cinq ne sont plus disponibles]

[7] WILL, Leonard D. Comparison of thesaurus management software for Pcs. Table of features, site Willpower Information, dernière mise à jour 06-09-2006. www.willpower.demon.co.uk/thessoft.htm [Présentation succincte de 45 logiciels thésaurus mise à jour en 2006. La bible ... anglo-saxonne. Sur le même site (-/thestabl.html), tableau mis à jour en 2004 des fonctionnalités de sept de ces logiciels]

[8] « Thesaurus management software ». *In : Encyclopedia of Library and Information Science*, vol. 51, suppl. 14, p. 389-407, 1993 [Spécifications pour des modules indépendants, expliquées indépendamment des produits du marché]

Thésaurus de recherche

[9] ERLOS, Frédéric. « Thésaurus et accès à l'information. Référentiels terminologiques adaptables au contexte : l'exemple d'un système de recherche d'informations dans une grande entreprise ». *7es Journées internationales d'Analyse statistique des Données Textuelles (JADT)*. www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/toc)ADT2004.htm [À caractère scientifique]

[10] FEYLER, François. « De la différence entre langage d'indexation et langage d'interrogation ». *SavoirsCDI*, mars 1999. http://savoirscdi.cndp.fr/culturepro/actualisation/linguistique/langagefeyler.htm

[11] LEFÈVRE, Philippe. *La recherche d'informations : du texte intégral au thésaurus*. Paris : Hermès, 2000. 253 p.

[12] MENON, Bruno. « Les nouveaux (?) usages du thésaurus pour l'accès à l'information ». Communication faite à l'ADBS, 5 février 2003. www.bmenon.net/

[13] SHIRI, Ali Asghar, REVIE, Crawford. « Thesauri on the Web: current developments and trends ». *Online Information Review*, 2000, vol. 24, n° 4, p. 273-279

[14] UK Government, CabinetOffice. *Design/selection criteria for software used to handle controlled vocabularies* (version 1.4 du 29/8/2006), www.govtalk.gov.uk/interoperability/gcl_document.asp?docnum=954 [Fonctionnalités à prendre en compte par les concepteurs d'application pour l'utilisation (indexation, consultation et recherche) d'un vocabulaire contrôlé, mais également pour sa gestion. Dans le cadre de l'administration électronique britannique]

[15] *Zthes specifications for thesaurus representation, access and navigation*. http://zthes.z3950.org/

Vocabulaires contrôlés et terminologie

[16] DAVIES, Ron . « Les nouvelles tendances des langages documentaires ». *Cahiers de la documentation*, mars 2006, 60e année, n° 1, p. 4-10

[17]]UN, Wang . « A knowledge network constructed by integrating classification, thesaurus and metadata in a digital library ». *Bulletin of ASIST*, December 2002-January 2003, vol 29, n° 2, p. 24-28

[18] Norme ISO 16642 (TMF - Terminological Mark-up Framework). TermsSciences : www.termsciences.fr/article.php3?id_article=18