

Quelles relations entretiennent les thésaurus documentaires et l'informatique documentaire ? Cette question est abordée par Sylvie Dalbin, sous l'angle technique des fonctionnalités et des usages, dans une autre contribution à ce numéro. Dans le présent article, elle se place dans une perspective historique pour étudier un demi-siècle de relations marquées par l'apparition des serveurs de banques de données professionnelles, par les évolutions induites dans les centres documentaires, dans les entreprises et dans les organisations, puis par l'apparition et le développement du Web.

par SYLVIE DALBIN

Thésaurus et informatique documentaires

Des Noces d'Or

■ PENDANT LES ANNÉES CINQUANTE¹, pour résoudre la « crise de la documentation scientifique », de grandes banques de références documentaires sont créées, qui mettent à disposition des fonds d'articles de revues, de rapports techniques, de communications de congrès, etc. Ces banques de données prennent appui sur les capacités des premiers ordinateurs et sont rapidement associées à des serveurs spécialisés qui en proposent l'accès à travers les réseaux de télécommunication, d'abord aux États-Unis avec Dialog/Lockeed et Orbit (On line Retrieval of Bibliographic Information Timeshared) au début des années soixante, puis partout à travers le monde, notamment en France avec Questel en 1978.

Il apparut bientôt que les index produits pour répondre à des usages « papier » (KWOC, KWIC) ne suffisaient plus à répondre aux besoins. Le traitement documentaire à partir de *descripteurs*, puis de *thésaurus documentaires*² [voir hors texte page 78] apportait une réponse concrète et efficace pour un accès à distance à cette documentation. Ces thésaurus sont donc apparus dès l'origine de l'informatique, grâce aux possibilités conjointes de cette technique et de la logique combinatoire.

La recherche documentaire en ligne associée aux thésaurus documentaires a donc à présent cinquante ans.

Parmi ces thésaurus pionniers, citons le MeSH (1964) associé à la base Medlars et établi à partir de l'*Index Medicus* (index d'articles publié entre 1879 et 2004) ou le *Chemical Engineering Thesaurus* réalisé par Mortimer Taube (1961). En France, le système DARC (Description, Acquisition, Recherche et Corrélation) dédié à la chimie fut développé par Jean-Émile Dubois à partir de 1954. Ce fut une époque d'intense « production documentaire » et de construction parallèle de thésaurus : il s'agissait de capitaliser les ressources scientifiques et techniques et de les rendre accessibles via les réseaux de télécommunication.

1 Les serveurs de banques de données professionnelles

Les serveurs proposent un accès unifié à un ensemble de bases documentaires, aux contenus variés (chimie, physique, nucléaire, médecine, etc.), dont les structures documentaires et les langages contrôlés sont adaptés à ces données, l'ensemble étant exploitable par un langage de commande informatique spécifique à chacun des serveurs. Aux difficultés inhérentes à la recherche multisources – que certains découvrent aujourd'hui sur le Web – s'ajoutaient alors des obstacles dus à la faiblesse des débits et aux coûts élevés des équipements nécessaires, obstacles freinant considérablement l'utilisation de ces fonds.

Les serveurs ont donc cherché en permanence à améliorer l'efficacité de leurs services, d'abord en enrichissant dès les années quatre-vingt les fonctionnalités proposées : recherche sur index groupés (index de base), recherche multibases (OneSearch de Dialog ou Duplicate de STN ; DialIndex de Dialog qui fonctionne comme les métamoteurs actuels), tri des résultats, etc. Parmi ces fonction-

nalités d'orientation technique, citons une fonction dédiée aux langages documentaires, *zoom*, qui oriente l'interrogateur vers les sources les plus adaptées à partir de l'analyse des termes contrôlés recueillis au cours d'une première recherche et affichés par ordre décroissant d'occurrence et par bases. Toujours à cette période, l'interrogation *plein texte* sur les notices est proposée avec de nombreux opérateurs dont ceux de proximité. Les ressources des micro-ordinateurs ont été mises à contribution dès l'apparition de ces outils, et les offres sur cédérom se sont multipliées au début des années quatre-vingt. En France, le Minitel, souvent émulé sur un ordinateur pour bénéficier des possibilités de stockage, a été un vecteur de démocratisation de l'accès à ces ressources documentaires.

L'intérêt porté à ces ressources d'information par des publics spécialistes des domaines considérés mais non-spécialistes de la recherche documentaire a conduit les serveurs à développer différentes offres. Des interfaces guidées, accompagnées d'une structuration par grands domaines des dizaines de banques de données proposées, offraient d'honnêtes solutions aux non-professionnels de la recherche documentaire, mais aussi aux spécialistes occasionnels. Des interfaces hors ligne – STN Express de STN ou Imagination de Questel-Orbit – permettaient aux spécialistes visés, sans contrainte de coût de connexion, de préparer des requêtes précises et de les sauvegarder. Dans toute cette effervescence, les vocabulaires contrôlés, en particulier les thésaurus, sont restés sous leur forme papier pour l'interrogation de ces bases professionnelles.

Avec l'arrivée du Web au début des années quatre-vingt-dix, les serveurs, cherchant à étendre leur clientèle aux internautes, ont poursuivi leurs efforts sur ce nouveau réseau qu'est l'internet. Des interfaces similaires à celles des moteurs de recherche furent proposées : simplifiées à l'extrême, avec toutefois la possibilité d'utiliser des commandes associées à des opérateurs plus complexes et plus efficaces.

Puis plus récemment (2000), et parallèlement à une concentration du marché des serveurs, le formidable développement de l'informatisation des activités et de la production de documents numériques a conduit à deux stratégies distinctes suivant les familles de données diffusées.

Les serveurs « presse » ou proposant des informations à haute teneur textuelle, fortement numérisées comme pour la documentation juridique³, ►

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

¹ Pour la période précédant les années cinquante, en particulier sur le développement des centres documentaires spécialisés, voir [3].

² Rappelons que 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).

³ Le domaine juridique a une ancienne pratique de la documentation numérisée. Voir [6] ainsi que sur le site de Stéphane Cottin : *Historique de la documentation juridique « électronique »*, 31 mai 2003, www.servicedoc.info/Historique-de-la-documentation.html.

Aux origines des termes descriptor, thesaurus et information retrieval

Le principe du *descriptor** est posé dès les années cinquante par Calvin Northrup Mooers**, du Massachusetts Institute of Technology (MIT), dans le cadre de solutions mécanisées. Celui-ci introduit les termes *descriptor* et *information retrieval* dans sa thèse de 1949. Ces premières réalisations ont conduit au thesaurus comme dictionnaire de descripteurs.

Le terme même de *thesaurus* appliqué en recherche documentaire (*information retrieval*) est souvent attaché au nom de Peter Luhn d'IBM (1957). Mais avec Luhn le thesaurus est associé à des traitements automatiques statistiques. Nous pensons plus vraisemblable d'inscrire la notion classique de thesaurus documentaire (utilisé pour l'indexation humaine) dans la descendance des travaux d'Helen Louise Brownson, de l'American National Science Foundation (ANSF). Lors d'une intervention faite à la « Dorking conference on classification research », pendant

cette même période (1957), celle qui avait précédemment été secrétaire de Vannevar Bush, le père de l'hypertexte, parlait en effet d'« *application of a mechanized thesaurus based on networks of related meanings* ».

* « *Descripteur : mot-symbole ou groupe de mots, représentant une idée ou un concept, généralement de portée assez large* », utilisé par Mooers dès 1947 (voir : www.cbi.umn.edu/collections/inv/cbi00081.html et <http://web.utk.edu/~alawren5/mooers.html>). Le système UNITERM développé par Mortimer Taube (IBM) s'appuyait sur les propositions de Mooers [1].

** Voir : Eugene Garfield. « A tribute to Calvin N. Mooers, a pioneer of information retrieval ». *The Scientist*, 1997, vol. 11, n° 4, p. 9. [http://garfield.library.upenn.edu/commentaries/tsv11\(06\)p09y19970317.pdf](http://garfield.library.upenn.edu/commentaries/tsv11(06)p09y19970317.pdf)

ont pu passer rapidement à des techniques d'accès « texte intégral » et de diffusion numérique des documents, avec une politique commerciale et technique de mise à disposition directe sur les intranets. Les traitements des corpus éditoriaux se sont fortement automatisés tant sur le plan de l'indexation⁴ que sur celui de leur classification, cette dernière pouvant être soit totalement automatique, soit supervisée à partir de vocabulaires contrôlés existants. Fréquemment, une indexation humaine, plus légère, complète ces traitements informatisés.

En même temps que cette numérisation (au sens de production d'information numérique), un important travail de structuration⁵ des documents et des métadonnées a été mené. Pour d'autres types de secteurs où le développement du document numérique s'avérait plus difficile⁶, le modèle d'accès à l'information est resté le même, les serveurs facilitant l'acquisition des documents directement à partir de la notice.

Par ailleurs, la démocratisation de la publication et de la diffusion électroniques a eu pour effet de pousser certains producteurs de bases documentaires à devenir leurs propres diffuseurs et à innover à partir d'une réflexion sur les usages de leurs publics et les spécificités de leurs fonds. On peut citer l'interrogation des bases brevets en langage naturel via la classification des brevets avec Lingway⁷ ou encore l'interrogation conjointe d'Embase et Excerpta Medica. Les langages contrôlés gardent ici leur rôle d'accès par sujets, mais le plus souvent en *back office* comme pour la fonction d'expansion du MeSH sur PubMed [3]. Notons que le développement d'interfaces dédiées à des fonds particuliers est une constante chez les serveurs⁸. Dans tous les cas, les ressources continuent à être indexées avec des thesaurus et un grand nombre de nomenclatures spécialisées.

2 Dans les centres documentaires

Parallèlement à ce mouvement lié aux banques de données professionnelles et au même moment, de nombreux changements ont lieu dans les centres documentaires et les bibliothèques spécialisées.

Les années soixante-dix voient la multiplication des bases documentaires dans les organismes de recherche et les grandes entreprises, avec à la clé la construction de thesaurus spécialisés. Ce qui est accessible sur les réseaux informatiques des entreprises reste la notice avec les indexats⁹. Le thesaurus lui-même est encore utilisé sur support papier sous ses différentes formes de présentation : liste alphabétique (globale avec ou sans environnement sémantique), liste permutée, champs sémantiques

et aussi schémas fléchés. La conception et la maintenance de ces thésaurus font souvent l'objet d'applications internes. Mais cette période voit aussi se développer des logiciels de gestion documentaire et des modules « thésaurus » associés, comme ceux de Mistral (Bull), de Basis (Open Text) ou de Texto.

La micro-informatique, au tout début des années quatre-vingt, a donné une forte impulsion aux bases documentaires et aux thésaurus. L'informatique sur micro-ordinateur offre alors aux documentalistes la possibilité de rechercher et de sélectionner des descripteurs au sein du thésaurus, celui-ci étant « embarqué » dans le logiciel. Le modèle indexation-thésaurus-recherche est alors conforté et s'installe pour les vingt années suivantes avec l'apparition de modules dédiés utilisables d'une façon identique pour la recherche et pour l'indexation : Polybase de Polyphot, JLBDoc, le module Thesaplus proposé avec Texto ou encore Superdoc sont ainsi disponibles dès 1982.

Mais l'informatique de l'époque a poussé les développeurs à opérer certaines simplifications fonctionnelles : ainsi de la polyhiérarchie, des fonctions de renvois d'une notion vers deux descripteurs ou encore des schémas fléchés¹⁰, pour ne citer que trois fonctions expressément décrites dans les normes des années soixante-dix. Mises en œuvre manuellement dans une période plus ancienne, elles sont quasiment absentes des modules « thésaurus » des logiciels documentaires de cette période. Entre 1990 et 1998, la GEIDE¹¹ n'a pas modifié ce schéma global. Durant vingt ans, une partie importante de la profession a fini par adopter la vision des thésau-

rus telle qu'elle s'est construite dans ces centres documentaires et bibliothèques spécialisées.

3 Dans les entreprises et les organismes

Parallèlement au déploiement des bases documentaires dans les centres documentaires des organismes, les « technologies de l'information » étaient utilisées pour informatiser de nombreuses activités professionnelles, en premier lieu la production des documents et plus généralement de l'information et des données. Cette informatisation s'est opérée dans les organismes à des rythmes différents suivant les secteurs professionnels et les domaines, mais de façon continue. Vivier de nouvelles pratiques liées à l'information et aux documents numériques dans le cadre de systèmes de veille, de gestion de connaissances et plus récemment de *records management*, cette « informatique documentaire » s'est très tôt orientée vers le « texte intégral » et plus récemment vers l'articulation de ces techniques automatiques avec des nomenclatures métiers.

Face à des flux et des volumes toujours plus importants, se trouve ici cristallisée la double problématique documentaire de fédération de ressources multiples et variées et d'exploitation du contenu d'importants corpus numériques. Les vocabulaires contrôlés ou taxonomies, pour utiliser le vocabulaire de ces environnements professionnels, sont plus que jamais présents, sous diverses formes : le thésaurus dans sa version simplifiée en liste de synonymes, de multiples nomenclatures associées à des métadonnées métiers, des classifications pour les gestionnaires de contenus et pour les portails, un outillage évolué avec des moteurs linguistiques, ou encore des ontologies pour des dispositifs de gestion de connaissances. Un certain nombre de rachats – Askonze par Documentum en 2004 pour la fédération de ressources, Datops en 2006 par LexisNexis pour l'exploitation de contenu, et Synapse par Factiva en 2005 pour la gestion des thésaurus et des taxonomies¹² – constituent des indices forts de cette situation.

4 Et le Web ?

À partir des années quatre-vingt-dix, la technologie du Web a impulsé un formidable mouvement autour des documents numériques, de l'accès à l'information via les réseaux informatiques, mais ►

⁴ Par exemple, CedromSNI utilise depuis 2001 les technologies d'analyse linguistique du canadien Nstein. Voir aussi : Séminaire Fédérer : « L'intégration des contenus Presse sur Intranet », GFII, 21 octobre 2003, www.gfii.asso.fr/article.php3?id_article=1339.

⁵ LexisNexis propose à l'utilisateur quarante champs de recherche [10].

⁶ En raison de la taille des documents, d'une chaîne de production encore fortement papier ou plus souvent des freins humains face à tous ces bouleversements.

⁷ Voir : Sabine Darrigade, Michèle Lyon-Bougeat et Bernard Marx. « Accès aux brevets en langage naturel Le système CIB-LN de l'INPI ». *Documentaliste - Sciences de l'information*, juin 2001, vol. 38, n° 2, p. 100-110.

⁸ Par exemple, des interfaces dédiées aux biographies ou aux dépêches de presse, chez Pressed, ou la recherche sur des marques chez Dialog [10].

⁹ Indexat : ensemble des termes ou indices issus de l'indexation d'un document et associés à ce document. Autre terme utilisé : *formule d'indexation* (Jacques Maniez), terme utilisé avec une autre signification dans d'autres secteurs.

¹⁰ Schémas fléchés dont la logique d'usage revient en force avec les cartographies sur le Web.

¹¹ GEIDE (gestion électronique d'information et documents existants) : association d'un document à la notice, le document étant sous une forme à l'époque non exploitable informatiquement, pour des raisons organisationnelles (chaîne de récupération et numérisation) et de droit (le document étant à l'origine externe à l'entreprise).

¹² Factiva Synaptica Knowledge Management System, sur le site de Factiva (www.factiva.com).

Ressources bibliographiques

Histoire de la documentation ou des sciences de l'information

- [1] *Chronology of information science and technology*. Website originally developed by Laird Whitmire, 1997; edited and revised by Lisa Gieskes, Spring 2002. www.libsci.sc.edu/BOB/istchron/ISCNET/ISCHRON.HTM
- [2] *Pioneers of information science in North America*. A project of SIG/HFIS (History and Foundations of Information Science), American Society of Information Scientists (ASIS). [1999]. www.libsci.sc.edu/Bob/ISP/isp.htm
- [3] FAYET-SCRIBE, Sylvie. *Histoire de la documentation en France : culture, science et technologie de l'information. 1895-1937*. Paris : CNRS Éditions, 2000. 313 p. (CNRS Histoire)
- [4] FONDIN, Hubert. « La science de l'information ou le poids de l'histoire ». In : *Les enjeux de l'information et de la communication*, Gresec, 24 mars 2006. http://w3.u-grenoble3.fr/les_enjeux/2005/Fondin/index.php (ou http://w3.u-grenoble3.fr/les_enjeux/2005/Fondin/fondin.pdf)
- [5] COMBEROUSSE, Martine. *Histoire de l'information scientifique et technique*. Paris : Nathan Université, 1999. (Collection 128)
- [6] FROCHOT, Didier. *Histoire des bases de données juridiques en France. 1 : Les origines*. Novembre 2005. www.les-infostrategies.com/article/0511109/histoire-des-bases-de-donnees-juridiques-en-france-1-les-origines
- [7] SERRES, Alexandre. *Chronologie générale de la documentation, d'hypertexte et d'Internet*. Urfist de Rennes, 2003. www.uhb.fr/urfist/HistInt/

Banques de données professionnelles

- [8] LARDY, Jean-Pierre. *Les accès électroniques à l'information : état de l'offre*. Paris : ADBS Éditions, 1993. 90 p. (Sciences de l'information. Série Recherches et documents)
- [9] LARDY, Jean-Pierre. « Savoir évaluer une interface ». In : *Acquérir et exploiter ses bases de données en ligne : guide pratique*. Paris : IDP, 2004. (Archimag Guide pratique) [Sur les fonctionnalités des interfaces et les langages de commandes.]
- [10] FOENIX-RIOU, Béatrice. *Recherche et veille sur le Web visible et invisible : agents intelligents, annuaires sélectifs, interfaces des grands serveurs, portails thématiques*. Paris : Bases Publications : Éditions Tec & Doc, 2001. 240 p.
- [11] MESGUICH, Véronique, THOMAS, Armelle. *Net recherche : le guide pratique pour mieux trouver l'information utile*. 2^e éd. augm. et mise à jour. Paris : ADBS Éditions, 2007. 159 p. (Sciences et techniques de l'information)

Vocabulaires contrôlés

- [12] VICKERY, Brian Campbell. « Thesaurus – a new word in documentation ». *Journal of Documentation*, December 1960, vol. 16, n° 4, p. 181-189
- [13] HJØRLAND Birger. *Lifeboat for KO. Specific systems (Classification systems, thesauri, etc.)*. www.db.dk/bh/Lifeboat_KO/SPECIFIC%20SYSTEMS/specific_systems.htm [Approche historique]

aussi autour de l'idée d'une informatique plus simple, plus souple, plus riche.

Dans les centres documentaires et les serveurs de banques de données professionnelles dès 1990, un peu plus tardivement dans les bibliothèques, le Web a été utilisé comme un autre réseau de télécommunication pour la diffusion des banques de références documentaires ou des catalogues, à la suite de Transpac et du Minitel. Après une période, courte mais douloureuse, entre 1993 et 2000, du tout automatique (indexation, classification) où les principes mêmes des bases documentaires et bien sûr des thésaurus étaient remis en cause, on est revenu depuis quelques années à des positions plus modérées mais aussi plus crédibles : il s'agit désormais, au lieu de les mettre systématiquement dos à dos¹³, de prendre en compte le meilleur des différents modèles ou techniques.

Il nous semble que l'on assiste aujourd'hui à une certaine convergence des problématiques documentaires entre ces trois environnements – organismes, lieux documentaires dédiés (archives, bibliothèques, centres documentaires) et serveurs de banques de données professionnelles. Chacun avait abordé et traité la « documentation informatisée » selon des méthodes et avec des outils différents, traçant durant ces cinquante dernières années des histoires relativement autonomes.

Mais aujourd'hui chacun prend en charge des documents numériques articulés à tout un ensemble de métadonnées construites pour partie à partir de vocabulaires contrôlés ; les interfaces orientées utilisateurs, ergonomiques, prennent toutes en compte les règles de l'art du Web ; et les outillages informatiques structurés optimisent et simplifient la production et l'accès à l'information.

Si un consensus semble se dessiner sur la nécessité d'innover en exploitant au mieux les techniques et méthodes les plus nouvelles, la tendance est encore grande de s'appuyer sur des pratiques anciennes pour définir les architectures des systèmes d'information. Ainsi du document qui a du mal à sortir de son schéma traditionnel ou des thésaurus coincés entre l'indexation et la recherche. Et que dire de l'utilisateur, confiné depuis les origines – image d'Épinal – dans le rôle de destinataire de systèmes mis en œuvre pour lui ?

Notion de document et place des utilisateurs : voilà deux points qui devraient structurer nos réflexions sur la place à venir des vocabulaires contrôlés et plus particulièrement des thésaurus.

¹³ En 1989 déjà, d'aucuns préconisaient cette solution. Voir : Ghislaine Chartron, Sylvie Dalbin, Marie-Gaëlle Monteil et Monique Vérillon, « Indexation manuelle et indexation automatique : dépasser les oppositions », *Documentaliste – Sciences de l'information*, juillet 1989, vol. 26, n° 4-5, p. 181-187.