THÉSAURUS

Étymologie et définition générale

- grec ancien « trésor, magasin.
- 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 (Documentation 1985).
- Thesaurus de linguistique, de médecine; thesaurus documentaire en sociologie; thesaurus de l'armement.
- Au moment de l'indexation des documents, les mots-clés sont choisis d'après un thesaurus hiérarchisé (...) qui fait l'objet de refontes successives.
- Selon Sylvie Dalbin, on parle 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 ».

Éléments d'histoire

- Selon Menon, on passe des grandes classifications encyclopédiques de bibliothèques, résultant du foisonnement éditorial de la fin du 19e siècle, aux listes de vedettes matières, qui « apparaissent et prospèrent à la faveur de l'accélération de la publication et de la circulation des savoirs sous forme imprimée, à l'aube du XXe siècle », jusqu'aux thesaurus, dont le développement a suivi la diffusion des techniques informatiques, vers la fin des années 1950.
- En premier lieu, il s'agissait de concevoir un système compact, qui ne demande pas de répertorier a priori de nombreux concepts, mais qui autorise l'expression d'un grand nombre de sujets.
- De nouvelles formes d'indexation, dites (post)coordonnées, sont mises au point.
- En deuxième lieu, il importait, en prenant en compte la synonymie, d'harmoniser le vocabulaire des auteurs, celui des indexeurs et celui des utilisateurs.
- Enfin, il fallait imaginer des moyens de guider indexeurs et utilisateurs dans le choix des termes appropriés, ce que vise la structuration sémantique des termes par les liens hiérarchiques et associatifs.
- On arrive ainsi aux trois caractéristiques définitoires du thésaurus : un langage structuré, contrôlé et combinatoire.
- 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 thésaurus est associé à des traitements automatiques statistiques.
- Nous pensons plus vraisemblable d'inscrire la notion classique de thésaurus 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 ».
- Ainsi, les thésaurus apparaissent après les classifications et sont associés au développement de l'informatique qui « rend désormais possible les manipulations automatisées et combinatoires du langage », en lien avec des logiciels documentaires.

Définition :

- Selon l'Association Française de Normalisation (AFNOR), un thésaurus est «un 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 les notions par des signes conventionnels. »
- Pour le thésaurus, il existe des normes nationales AFNOR Z 47 - 100 et internationales ISO 2788.

Caractéristiques principales

- Notion de descripteur : éliminer synonymie et polysémie Un concept = un descripteur et un seul Un descripteur = un concept et un seul

- Concevoir un langage univoque et post-coordonné: l'aspect combinatoire (associer les descripteurs pour l'interrogation) permet notamment avec les opérateurs booléens, de relier les éléments d'une indexation préalable
- Deux tâches distinctes avec un thésaurus : indexation et recherche
- Indexation : rigueur du choix des termes pour l'indexation ; le thésaurus comme un dictionnaire, utilisé pour contrôler ou affiner le choix d'un mot ou d'une expression
- Recherche d'information : contexte plus complexe pour les utilisateurs

Exemples de thésaurus

- Motbis
- Thesaurus de l'UNESCO
- Thésaurus de la formation

Au-delà des thésaurus

- Selon Chichereau et al. (2007)., « les thésaurus utilisés pour l'indexation et pour la recherche par mots clés se sont ainsi placés au centre d'un débat : les normes d'élaboration et de gestion des thésaurus datent, pour la plupart, des années quatre-vingt et accusent leur âge. D'autres outils – moteurs de recherche en texte intégral, taxonomies, ontologies – occupent, à tour de rôle, le devant de la scène dans le monde de la documentation. »
- Citons également les outils linguistiques, le développement de folksonomie avec la manipulation des signets, les ontologies...

RÉPERTOIRE DE VEDETTES-MATIÈRE

- Liste constituant un répertoire alphabétique des concepts utilisés (concepts sont représenté par un mot ou une expression)
- Langage précoordonné composé de vocabulaire de termes liés entre eux et d'une syntaxe qui indique les règles de construction de l'indexation.
- En France, sa première utilisation remonte à 1976 par la BPI (Bibliothèque Publique d'Information) qui décide d'utiliser le langage d'indexation de l'université de Laval.
- La BnF utilisera aussi la liste vedette matière, mais Rameau.
- Autre liste vedette matière : Blanc Montmayer.

INDEX

Définition

Généralités :

- D'après l'ADBS : « liste ordonnée de noms de personnes (index auteurs, index personnalités), de lieux (index des lieux), de matières (index des matières), etc., figurant dans un document ou dans un instrument de recherche (informatisé ou manuel), assortis d'une référence permettant de retrouver l'information. Les termes de cette liste constituent des points d'accès, permettant un accès différent de celui du document ou de la collection indexé(e) ».
  ➔ Un index peut se présenter sous la forme d'un document autonome, d'une annexe à un document ou à une partie de document, ou encore d'un fichier consultable en ligne sur un système informatique.

Index et Internet :

- En ce qui concerne l'internet et plus particulièrement les bases de données, un index est un élément qui permet d'optimiser les requêtes pour chercher et acquérir les informations sur la base de donnée, de la même manière qu'un index sur un livre.
  ➔ Répertorie les occurrences d'un mot.

Approche historique

- Dès l'Antiquité, l'alphabet a servi à établir des listes alphabétiques de noms.
- Les techniques de classement des entrées et de repérage restent balbutiantes jusqu'à l'invention de l'imprimerie.
  MAIS, apparition d'outils pour faciliter la lecture des livres pendant cette longue période et qui ont contribué à favoriser la pratique des index : les concordances, les glose, etc.
- Apparition des glossaires qui ont fortement influencé la naissance des index.
  ➔ Deux exemples sont devenus fameux : le glossaire de Reichenau (8ème siècle) et le dictionnaire de Papias (moine italien du 11ème siècle).
- Au 13è siècle, de façon exceptionnelle, existent déjà des spécimens d'index : index de concordance biblique, de citations et de sujets, dans la littérature hébraïque notamment.
- Au 14è siècle, les index deviennent d'usage courant dans les bibliothèques des monastères et des universités.
- Il faut attendre le 20è siècle pour que les techniques de production d'index se normalisent.
  ➔ La norme ISO 999.
  ➔ Les index ont plus évolué dans les deux dernières décennies que durant les six siècles précédents.

Enjeux actuels

- Dans les années 1970, les facilités de l'indexation automatique du texte intégral, rapide et peu coûteuse, ont généralisé le mode de recherche plein texte.
  ➔ D'abord avec de grosses bases de données textuelles, puis vingt ans plus tard avec des moteurs de recherche du web.
- Cependant, la machine a cependant ses limites.
  ➔ On n'a pas trouvé encore plus efficace que l'indexeur professionnel pour réaliser l'indexation matière tant les décisions à prendre impliquent une multitude de facteurs irréductibles à un traitement logique.
  ➔ Il faut parfois un peu de temps pour qu'un site soit indexé. C'est la raison pour laquelle les nouveaux sites ou nouvelles pages ne figurent pas sur les SERP immédiatement.

- L'indexation peut-être analytique ou automatique. Elle représente un enjeu majeur de la documentation (cf. les langages documentaires et l'analyse documentaire).