

2. L'indexation aujourd'hui

Muriel Amar

CONSERVATEUR DE BIBLIOTHÈQUES
BIBLIOTHÈQUE PUBLIQUE D'INFORMATION, SERVICE ÉTUDES ET RECHERCHE

Si les méthodes d'indexation, linguistique ou structurelle, bouillonnent avec l'expansion du Web, elles ne sont pas encore parfaitement adaptées aux besoins des utilisateurs. Un retour aux sources s'impose pour penser non plus en termes de « langage » mais de « discours » documentaire, prenant en compte le contexte thématique du document.



Site de la BPI: <http://www.bpi.fr>

1. Les intitulés de deux récentes journées d'étude sont à ce titre exemplaires: « La fin du catalogage! ? », proposée par Médiadix le 21 octobre 2004 et « L'indexation à l'heure du numérique » programmée par l'ADBS, avec beaucoup plus d'espoir, le 5 octobre 2004.
2. Une chaîne de caractères se définit comme

une suite de caractères comprise entre deux espaces. Or certains mots comprennent plusieurs chaînes de caractères, comme « pomme de terre »; certaines chaînes comprennent des signes typographiques séparateurs de mots comme dans « j'aime ». Pour une revue des problèmes, voir [FUC 1993].

A lors que la fin – voire la mort – du catalogage est régulièrement proclamée, l'indexation, elle, semble jouir d'une étonnante vitalité¹: au centre des préoccupations de plusieurs communautés de chercheurs et d'industriels, elle ne cesse d'apparaître dotée de nouveaux qualificatifs qui la rendent obscurément attractive. Tour à tour linguistique, conceptuelle, structurelle, l'indexation reste encore au-devant de la scène dans le très actuel projet du Web sémantique, grâce aux nouvelles formes de représentation que sont les ontologies.

Parallèlement, la pratique professionnelle de l'indexation, qu'elle porte sur le document imprimé ou sur le document numérique, se signale par une remarquable stabilité: corsetée par la norme Z 47-102 (1978) (voir [AFN 1978]), l'indexation reste une entreprise de contrôle terminologique qui peine à prendre en compte les caractéristiques du document numérique.

Reste que, aussi bien du côté des techniques – constamment renouvelées – que des pratiques – très peu bousculées – de l'indexation, les questions demeurent identiques: comment rendre compte des thèmes d'un document? À quoi conduisent les mots d'indexation?

Techniques d'indexation: un univers en expansion

Nous désignerons par « techniques d'indexation » l'ensemble des techniques informatiques qui manipulent les chaînes de caractères d'un document pour en donner accès.

Le postulat de base de ces techniques est que « les mots d'un texte ont une signification par eux-mêmes » et que « le mot d'un texte est lui-même un index. Tous les moteurs de recherche reposent sur ce principe-là: un texte est sa propre indexation, à un niveau zéro » [BAC 1999].

Le problème central à résoudre est celui de la non-coïncidence entre « chaîne de caractères » et « mot² ».

Deux principaux types de communautés explorent cette problématique: celle de l'ingénierie linguistique et celle de l'ingénierie des connaissances.

3. Minimale, ce type de logiciel procède à une analyse morphosyntaxique des textes; certains proposent des modules de « reformulation » et/ou d'expansion sémantique. Pour un panorama récent, voir [CHA 2003].

4. Ou, d'un point de vue strictement linguistique, des unités nominales référentielles. Sur le rapport terme et descripteur, voir [LEG 1984].

5. Voire aussi celles qui lui sont communément associées comme l'encéphalopathie spongiforme bovine, par exemple.

6. Sur le thème de la « vache folle », l'utilisation du thésaurus du *Monde*, par exemple, obligerait à recourir à la combinaison de descripteurs suivante: maladie animale, viande, bétail.

7. Voir, par exemple, l'application *Cour des comptes*: www.ccomptes.fr/recherche/recherche.htm ou celle des AGF relatée dans [DAL 2000].

8. Voir, par exemple, [RIC 2002].

9. Pour une revue complète de l'indexation conceptuelle et structurelle, avec des exemples de contextes d'application, voir [IND 2000].

10. De ce point de vue, l'index n'est plus uniquement un index de contenu du document, il y a aussi des index rendant compte de la structure du document ou renvoyant à un cheminement, un parcours de lecture interne et/ou externe au document. Voir notamment [GER 2002].

11. En théorie. Le problème est connu de la variabilité, aussi bien dans le choix des thèmes que dans leur nom, que ce soit entre indexeurs ou pour un même indexeur dans la durée. Sur ce point, une synthèse récente dans [MOU 2002].

12. En l'absence du texte intégral du document primaire, sur quelle autre base que des descripteurs rendre des documents d'une part repérables par leur contenu et d'autre part commensurables, c'est-à-dire passibles d'un jugement de similarité ou de dissemblance thématique?

Partir de l'analyse du texte: l'indexation linguistique

Pour la communauté de l'ingénierie linguistique, issue de la linguistique formelle, il s'agit de transformer des chaînes de caractères informatiques, dépourvues de signification, en un ensemble de termes interprétables par des êtres

humains. Les modèles et les outils

développés proposent une analyse

linguistique des textes³ en iso-

lant, notamment pour les logi-

ciels d'extraction terminolo-

giques, l'ensemble des unités

nominales susceptibles de

désigner des objets du monde :

ces unités aux propriétés lin-

guistiques spécifiques se nom-

ment des termes⁴. Ce type d'in-

dexation dite linguistique permet ce que

Maniez appelle la « recherche contextuelle »

dans les documents textuels numériques [MAN

2002 : 91]: à partir d'une expression singulière,

d'un « terme » comme « vache folle », par

exemple, le système de recherche exploitant ces

technologies retrouve tous les contextes où est

employée l'expression saisie en requête⁵.

En explorant les « terminologies » à l'œuvre

dans les textes, ce type d'indexation fournit, le

plus souvent, des points d'accès aux documents

proches des formulations familières des utilis-

ateurs de systèmes d'information. C'est pour cette

« convivialité⁶ » que, dans certains contextes

documentaires et pour certains types de besoin

d'information, les outils issus de l'ingénierie lin-

guistique sont utilisés⁷.

Développer l'intelligence artificielle: l'indexation structurelle

Pour la communauté de l'ingénierie des connais-

sances, issue de l'intelligence artificielle, la pro-

blématique est légèrement différente: l'indexa-

tion des documents est orientée non plus vers

l'utilisateur final mais vers la machine: il s'agit

alors de « *fournir un marquage des contenus*

[...] interprétable par des machines rendant

ainsi possible l'automatisation de nombreuses

tâches aujourd'hui accomplies par des êtres

humains » [MEN 2004].

C'est ainsi que se sont développées les tech-

niques d'indexation conceptuelle et structurelle

aujourd'hui utilisées dans le cadre des travaux

du Web sémantique⁸, qui se veut une extension

du Web actuel et vise à rendre les contenus non

plus uniquement accessibles et affichables, mais

aussi exploitables et interprétables par des

machines. L'enjeu de l'indexation est alors d'ajou-

ter aux documents des « connaissances », le plus

souvent expertes, pouvant servir dans le cadre de

calculs, c'est-à-dire dans la réalisation de tâches

précises (le diagnostic de pannes dans un envi-

ronnement technique, par exemple⁹). Ces

connaissances rendent compte, sous forme

d'« annotations », d'une lecture et d'une inter-

prétation expertes des documents (c'est l'in-

dexation conceptuelle); ces annotations

sont contraintes par des schémas

d'annotation, par des « conte-

neurs de connaissances » (the-

saurus, terminologies, ontolo-

gies) et par l'appartenance

d'un document à un genre

(c'est l'indexation structu-

relle).

Dans ce cadre, l'indexation n'a

plus pour seul objectif de per-

mettre des recherches ultérieures;

elle a aussi celui d'exploiter et de

mobiliser au mieux, le moment voulu, les

informations nécessaires à la réalisation d'une

tâche bien précise. Cette approche revient à

considérer l'indexation comme « la documenta-

tion d'une tâche d'exploitation d'un document »

[IND 2000 : 11-35].

Vers une recherche thématique orientée utilisateur: l'indexation professionnelle

Si l'univers des techniques d'indexation est en

pleine expansion, notamment dans les commu-

nautés de l'ingénierie linguistique et de l'ingé-

nierie des connaissances, on voit bien pourquoi

ces techniques touchent si peu les pratiques des

professionnels de l'information: les orientations

« recherche contextuelle » pour l'indexation dite

linguistique, et les orientations « machine » pour

les indexations structurelle et conceptuelle sont

effectivement fort éloignées des préoccupations

de l'indexation professionnelle, destinée aux uti-

lisateur finaux désireux de mener une recherche

thématique.

Cependant, ces techniques d'indexation, parce

qu'elles ont pour objet premier (et exclusif) le

document numérique, soulignent des spécifici-

tés que doivent (ou devraient) prendre en compte

les professionnels de l'information.

On retiendra notamment que le passage du

support analogique au support numérique per-

met :

– de penser des « index » non plus distincts du

document primaire mais au contraire issus et/ou

intégrés à celui-ci: ces index sont non seulement

des outils de recherche mais aussi et surtout des

outils de lecture¹⁰, sous réserve qu'ils relèvent

du statut linguistique adéquat (unités nominales

référentielles) ;

– de manipuler non plus l'intégralité d'un docu-

ment mais aussi des segments pouvant, le cas

échéant, être recombinaés pour produire de nouveaux documents, sous réserve que soient introduites des connaissances contextuelles, externes aux documents.

Pratiques d'indexation : les limites de la continuité

Nous désignons par pratiques d'indexation toutes les pratiques professionnelles d'analyse de contenu se rapportant aux normes professionnelles et notamment à la norme Z 47-102.

Celles-ci ont un double objectif : identifier les thèmes d'un document et fournir à l'utilisateur des renseignements sur les « choses » et non sur des « mots », c'est-à-dire un ensemble de documents homogènes d'un point de vue thématique.

La réalisation de ces deux objectifs passe par le recours à un langage documentaire, qui donne la liste de tous les thèmes identifiables dans un fonds, et le nom à utiliser pour rendre compte des thèmes sélectionnés en suivant la règle dite d'univocité : à un même thème identifié, l'indexeur donne toujours le même nom¹¹.

Si ces pratiques d'indexation contrôlée ont fait leur preuve pour le support imprimé¹², que deviennent-elles dans le cas du document numérique ?

Un nécessaire retour aux sources

La thématisation réalisée à l'aide d'un langage documentaire est nécessairement partielle (ou sélective) et prédéfinie (les thèmes autorisés sont déjà connus). Cette monothématisation prédéfinie¹³ ne permet pas de donner accès au « texte intégral » du document, c'est-à-dire aux thématiques multiples dont il est porteur¹⁴; ou, plus exactement, à quoi bon disposer du texte intégral d'un document si c'est pour y avoir accès par une seule thématique, toujours la même ?

Par ailleurs, les clés d'accès – les descripteurs – retenues pour nommer les thèmes ne permettent de désigner des « objets¹⁵ » que s'ils sont restitués dans le cadre de leur réseau de relations hiérarchiques, associatives ou définitoires¹⁶. Or, il est très difficile d'exiger des utilisateurs finaux qu'ils maîtrisent les relations des langages documentaires : le plus souvent, ils utilisent les descripteurs comme de simples « mots » ; mais les mots seuls n'ont pas la possibilité, dans la langue, de désigner des objets [MIL 1989].

Une façon de ne pas sombrer dans le désespoir devant le constat d'inadéquation de l'indexation contrôlée dans le contexte numérique est de revenir aux fondements (linguistiques) de l'indexation : ne peut-on, autrement, réaliser les deux objectifs de l'indexation professionnelle, que sont la thématisation et la référencement ?



« ... si, par exemple, le mot *restauration* isolé dans le lexique français n'a pas de référence bien définie, le descripteur *restauration* en relation de spécificité avec le descripteur *architecture* permet, lui, de désigner un objet du monde. »

Fondements de la thématisation

D'un point de vue linguistique, la thématisation se réalise en deux étapes : une étape de construction du thème, qui est de nature discursive (et plus précisément interdiscursive, mettant en jeu plusieurs textes), et une étape de formulation, qui est de nature lexicale [MAR 1988 et 1997].

Dans l'indexation contrôlée, seul le résultat final – le choix du nom du thème – est donné aux utilisateurs. La construction du thème, qui se fait notamment en mobilisant des connaissances extérieures au document (connaissances de la discipline mais aussi connaissances de la collection déjà constituée), est réalisée par l'indexeur dans le secret de son *back-office*. Du coup, l'indexation contrôlée conduit à livrer aux utilisateurs une interprétation non documentée très difficilement reconstituable.

On voit bien tout l'avantage que l'indexation gagnerait à se situer, non plus en aval de la thématisation mais en amont, au niveau initial de la construction du thème. Dans ce cas, l'indexation consiste à maintenir les différents thèmes possibles, laissant ouverts, à l'utilisateur, tous les parcours interprétatifs : c'est donc ici lui et

13. Pour l'argumentation, voir [AMA 2000].

14. [FLU 1992 : 107] : « L'indexation, dans la mesure où elle transforme des notions trop précises en notions plus générales, empêche un accès aux textes par des questions trop pointues. L'indexation manuelle appauvrit la sémantique des documents en n'en donnant que les traits jugés, à un moment donné, essentiels ».

15. On désigne habituellement en linguistique par « référence » la « propriété d'un signe linguistique de renvoyer à un objet extralinguistique, qu'il soit réel ou imaginaire », [DIC 1994].

16. Les relations permettent, dans un langage documentaire, de stabiliser la référence des mots : si, par exemple, le mot *restauration* n'a pas de référence bien définie isolé dans le lexique français, le descripteur *restauration* en relation de spécificité avec le descripteur *architecture* permet, lui, de désigner un objet du monde.

non plus l'indexeur qui thématise, c'est-à-dire qui achève une lecture et la nomme. Indexer consiste alors à permettre la construction des unités d'interprétation que le texte propose, grâce à la mise en contexte du document¹⁷ : n'a-t-on pas alors la possibilité de donner accès au texte intégral des documents ou, plus exactement, à l'intégrité textuelle du document ?

On perçoit, sur ce point, ce que les techniques issues de l'ingénierie des connaissances pourraient nous apporter, par exemple, par des éléments de contextualisation des documents via des connaissances à la fois disciplinaires et bibliothéconomiques (relatives aux fonds documentaires déjà constitués).

Fondements de la référencement

Si on utilise des mots en indexation, ce n'est pas pour en donner le sens mais pour permettre de désigner des objets (processus de référencement). Or, d'un point de vue linguistique, toutes les unités ne sont pas référentielles. Pour la catégorie nominale, seuls les groupes nominaux le sont [MIL 1989].

L'indexation contrôlée repose sur ce paradoxe qui consiste à utiliser des unités nominales, non référentielles, pour référer de façon artificielle, par recours aux relations des langages documentaires : ne pourrait-on pas envisager d'utiliser le mode de construction référentielle « natu-

relle » des sujets parlants, les groupes nominaux et, en particulier, les unités terminologiques ?

Sur ce point, ce sont les résultats issus de l'ingénierie linguistique qui sont susceptibles de nous aider à identifier les clés d'accès les plus « naturelles » aux sujets parlants que sont les utilisateurs d'un système d'information.

« Si le rôle de construire et de nommer les thèmes revient alors à l'utilisateur, l'indexeur est celui qui le lui permet, grâce au discours documentaire et aux contextes qui rendent intelligibles et interprétables les thèmes d'un document. »

Du langage documentaire au discours documentaire
Face au double objectif de la thématisation et de la référencement, la pratique d'indexation se situe soit du côté du lexique soit du côté du discours.

Du côté du lexique, elle recourt au langage documentaire, chargé de la double fonction de nommer les thèmes et d'assurer la stabilité référentielle des unités lexicales. Mais cette indexation que l'on nommera lexicale peine à rester adéquate dans le contexte du document numérique : le filtrage thématique réalisé par le langage documentaire apparaît comme une restriction pénalisante et inutile, tandis que la stabilisation référentielle obtenue de force par les relations des langages documentaires contredit douloureusement le bon sens linguistique des utilisateurs.

Par opposition, apparaît un type d'indexation qui permettrait un accès direct au « texte intégral » des documents, en s'attachant à résoudre le double problème précédemment identifié : une construction référentielle naturelle aux

Références bibliographiques

[AFN 1978] Afnor. « Norme Z 47-102: principes généraux pour l'indexation des documents » in *Documentation*. 7^e éd. Tome I: présentation des publications, traitement documentaire et gestion des bibliothèques. Paris: Afnor, 2000. (Recueil de normes, règlements et certifications). P. 393-402

[AMA 2000] AMAR (Muriel) 2000. *Les fondements théoriques de l'indexation: une approche linguistique*, Paris: ADBS Éditions. (Sciences de l'information. Recherches et documents).

[AMA 2003] AMAR (Muriel) 2003. *Documentation et philosophie II. À propos de l'indexation discursive*. Textes réunis et présentés par Benoît Hufschmitt, Jean-Pierre Cotten et Marie-Madeleine Varet. Besançon: Presses universitaires franc-comtoises. (Annales littéraires de l'université de Franche-Comté. Série Philex; 7).

[BAC 1999] BACHIMONT (Bruno) 1999. *Atelier INA-Recherche*, n° 4, « Interfaces et outils d'analyse et d'indexation ». Séance du 21 juin 1999, compte rendu. www.ina.fr/inattheque/activites/ateliers/atelier4/A4_19990621.fr.html

[CHA 2003] CHAUMIER (Jacques) et Martine Dejean 2003. « Recherche et analyse de l'information textuelle: tendances des outils linguistiques ». *Documentaliste - Sciences de l'information*, vol. 40, n° 1, p. 14-24.

[DAL 2000] DALBIN (Sylvie) et Bruno Salléras 2000. « Une expérience d'utilisation d'un système d'information documentaire en langage naturel ». *Documentaliste - Sciences de l'information*, vol. 37, n° 5-6, p. 312-324.

[DIC 1994] *Dictionnaire de linguistique et des sciences du langage 1994*. Paris: Larousse.

[FLU 1992] FLUHR (Christian) 1992. « Le traitement du langage naturel dans la recherche d'information documentaire ». In *Les Interfaces intelligentes dans l'information scientifique et technique*, cours INRIA dirigé par Christian Bornes. Le Chesnay: INRIA. P. 105-128.

[FUC 1993] FUCHS (Catherine) sous la dir. de, 1993. *Linguistique et traitements automatiques des langues*. Paris: Hachette. (Supérieur).

[GER 2002] GERY (Mathias) 2002. « Un modèle d'hyperdocument en contexte pour la recherche d'information

17. Voir aussi [BAC 1999]: il s'agit de définir « une méthodologie d'indexation qui internalise au niveau du document des contextes de lecture extrinsèques au document. L'enjeu est de pouvoir traduire en termes de règles d'interprétation ce qui appartient au contexte dans le cadre d'une lecture endogène au document ».

18. Pour plus de détails sur la notion de discours documentaire, voir [AMA 2003].

sujets parlants et une thématisation non contrainte des documents. Nous nommerons ce second type d'indexation l'indexation discursive.

Dans le cadre de l'indexation discursive, l'attention n'est plus portée sur le lexique mais sur les discours et l'instrument privilégié n'est plus alors le langage documentaire mais le discours documentaire. Si le rôle de construire et de nommer les thèmes revient alors à l'utilisateur, l'indexeur est celui qui le lui permet, grâce au discours documentaire et aux contextes qui rendent intelligibles et interprétables les thèmes d'un document¹⁸.

Vers la prise en compte du contexte

Si c'est sous la forme de désespérantes parallèles que semblent se déployer les deux ensembles de techniques et de pratiques d'indexation que l'on voit aujourd'hui, il est possible d'y déceler des intersections fructueuses, pour peu que l'on veuille bien revenir aux fondements de l'indexation reformulés, sous un angle linguistique, par les deux termes de thématisation et de référenciation.

De ce point de vue, l'indexation se détechnicise pour redevenir une opération intellectuelle dont la vocation première touche la constitution même des collections et la maîtrise des fonds documentaires. En effet, la prise en compte des contextes, du « discours » dans l'opération d'indexation signifie que, avant de donner les « mots » pour dire les thèmes communs à plusieurs documents, on définit d'abord des ensembles thématiques : « *l'on va d'abord*



définir un document par un corpus qui permet de déterminer ses conditions de production et d'interprétation. Un document n'est pas un fait isolé. Il faut pouvoir le plonger dans le contexte empirique dans lequel il est attesté. On réintroduit la notion de contexte d'interprétation et d'énonciation qui avait été mise en avant par la pragmatique. On la réintroduit à un niveau documentaire, c'est-à-dire à un niveau presque philologique » [BAC 1999]. N'est-ce pas là un stimulant programme que peuvent se donner à suivre les professionnels de l'information ? ●

structurée sur le web ». In *Recherche et filtrage d'information*, sous la dir. de Catherine Berrut et Mohand Boughanem. Paris: Hermès. (*Ingénierie des systèmes d'information*, vol. 7, n° 1-2). P. 11-44.

[IND 2000] *L'Indexation*, sous la direction de Jean-Michel Jolion 2000. Paris: Hermès. (Document numérique, vol. 4, n° 1 – 2).

[LEG 1984] LE GUERN (Michel) 1984. « Les descripteurs d'un système documentaire: essai de définition ». *Condenser*, suppl. 1, p. 163-169.

[LEG 1991] LE GUERN (Michel) 1991. « Un analyseur morpho-syntaxique pour l'indexation automatique ». *Le Français moderne*, n° 1 (59), p. 22-35.

[MAN 2002] MANIEZ (Jacques) 2002. *Actualité des langages documentaires: fondements théoriques de la recherche d'information*. Paris: ADBS Éditions. (Sciences de l'information. Études et techniques).

[MAR 1988] MARANDIN (Jean-Marie) 1988. « À propos de la notion de thème de discours. Éléments d'analyse dans le récit ». *Langue française*, n° 78, p. 67-87.

[MAR 1997] MARANDIN (Jean-Marie) 1997. *Perception syntaxique et constructions syntaxiques. Mémoire d'habilitation*. Paris: Université Paris VII-Denis-Diderot.

[MEN 2004] MENON (Bruno) 2004. « Web sémantique et traitement automatique des langues ». In *Actes du colloque I-expo 2004*, intervention du 8 juin 2004, session Web sémantique: théorie et mise en œuvre. www.i-expo.net/documents/actes2004/i3/BrunoMenon.pdf.

[MIL 1989] MILNER (Jean-Claude) 1989. *Introduction à une science du langage*. Paris: Seuil. (Les Travaux).

[MOU 2002] MOUNIER (Évelyne) 2002. « Systèmes documentaires et systèmes de gestion de bibliothèques: place et rôle de l'opérateur professionnel ». In *Interaction homme-machine et recherche d'information*, sous la dir. de Céline Paganelli. Paris: Hermès; Lavoisier. (*Traité des sciences et techniques de l'information*). P. 103-132.

[RIC 2002] RICHY (Hélène) 2002. « Métadonnées et document numérique ». In *Les Techniques de l'ingénieur*. Paris: éditions des Techniques de l'ingénieur. Article n° H7155.