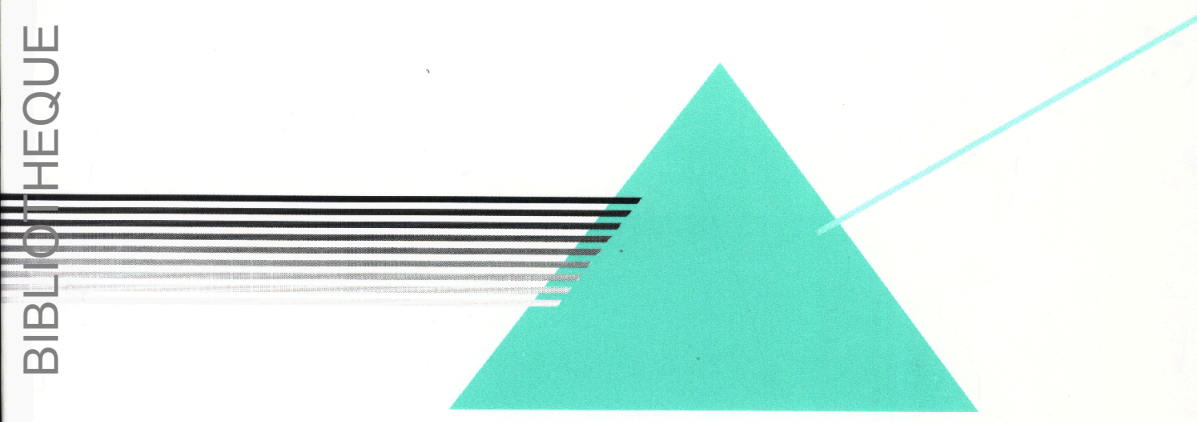
MURIEL AMAR

LES FONDEMENTS THÉORIQUES DE L'INDEXATION

UNE APPROCHE LINGUISTIQUE

BIBLIOTHEQUE DU CERIST



ADBS
ÉDITIONS

Muriel Amar

Les Fondements théoriques de l'indexation

Une approche linguistique

ADBS Editions

SOMMAIRE

PRÉFACE

Il y a au moins cinq cents ans que l'on indexe, et pourtant il aura fallu attendre le milieu du XX[e] siècle pour voir apparaître les mots *indexer* et *indexation*, que les dictionnaires datent de 1948. On a commencé par des ouvrages, et de là on est passé à des collections. Au début du XVI[e] siècle, ce n'est pas encore le mot *index* qui désigne le résultat de l'opération ; c'est *tabula*. Il s'agit bien de rompre la linéarité des documents traités, d'en donner une projection tabulaire, qui permette d'y tracer un chemin autre que celui que les auteurs avaient choisi. C'est ainsi, en partant de l'index, que la plupart des lecteurs se promenaient dans Pline l'Ancien, dans les *Essais* de Montaigne ou dans les *Commentaires hiéroglyphiques* de Pierius Valerianus.

En tête du tome V de ses *Diversités* (1610), Jean-Pierre Camus, l'évêque de Belley, ami de saint François de Sales, dit son hostilité à la pratique de l'indexation et au mode de lecture qu'elle induit. Il demande au lecteur de ne pas considérer comme une imperfection le fait que son livre soit « sans Indice des mémorables » : « C'est une erreur populaire, qui n'infecte que les faibles cerveaux, qui appellent cela l'âme du livre, et c'est l'instrument de leur stupidité. Ces gens peuvent être appelés *Doctores tabularii*, lesquels *sapiunt tantum per Indices*. Les enquerrez-vous de ce qu'ils savent ? Ils vous demandent un livre pour le montrer, et aussitôt à la Table pour trouver ce qu'ils cherchent, les habiles appellent cela le pont aux ânes. » Les quatre premiers volumes des *Diversités* étaient munis d'index, d'ailleurs fort bien faits, ce qui n'arrête pas les protestations de Jean-Pierre Camus : « Les tables des tomes précédents de l'auteur, faites par je ne sais qui, et à son insu, lui déplaisent, sachant qu'il faut retrancher tant que l'on peut ce qui fomente la paresse, paresse mère de l'ignorance. » Les volumes suivants comportent des index. Le fait que la protestation de Jean-Pierre Camus soit restée vaine, même auprès de ses propres éditeurs, montre que l'indexation répond à un véritable besoin, dès que l'imprimerie a multiplié les documents : on ne peut pas tout lire, de tous les livres, même en n'étant pas paresseux. À la nécessité empirique de trouver de l'information répond la pratique de l'indexation, qui restera empirique pendant plus de quatre siècles.

On a donc pu indexer, génération après génération, sans même éprouver la nécessité de nommer cette pratique, et *a fortiori* de la théoriser. Cela ne présentait pas d'inconvénient majeur dès lors que l'indexation était la tâche d'un homme seul : la qualité de l'indexation était fonction de la qualité de l'indexeur, et nombreuses sont les tables qui ne manquent ni de rigueur ni de cohérence. La limite de la masse de documents à indexer n'était limitée que par la puissance de travail de celui qui

s'en chargeait. On demandait à Du Cange, dont les glossaires sont sans doute la plus vaste entreprise d'indexation de l'époque classique, comment il avait pu mener à bien une telle tâche. C'était en y travaillant depuis l'âge de dix-huit ans, tous les jours, douze heures par jour, à une seule exception près : le jour de son mariage, il n'avait travaillé que huit heures.

La réduction des horaires et l'accroissement des collections condamnaient l'indexation à devenir une tâche collective, en posant de redoutables problèmes de cohérence. Il a donc fallu normaliser. Mais un savoir-faire empirique ne se laisse pas facilement normaliser, et les normes sans fondements sont les pires des ornières. Au mieux, tant qu'il ne s'agit que de coordonner des tâches qui restent purement humaines, on peut arriver à un semblant de cohérence, même si l'essentiel du consensus reste dans l'implicite. Mais, le jour où le recours aux moyens automatiques a imposé une rationalisation et une explicitation totale des procédures, force a été de constater qu'une approche empirique, même sous sa forme normalisée, ne suffisait plus. Il fallait donner à la vieille pratique de l'indexation, enfin nommée, des fondements théoriques.

Les premiers travaux sur l'indexation automatique ont cherché à faire simuler par la machine les procédures manuelles ; au mieux, on faisait moins bien. On continuait à penser dans le cadre d'un système documentaire qui répondait à une requête de l'utilisateur par une liste de références. Les progrès technologiques de ces dernières années ont tout bouleversé : la possibilité de transférer le document sur des supports informatiques a comme conséquence que l'utilisateur ne se contente plus de la référence ; il lui faut le texte lui-même. L'indexation manuelle était nécessairement partielle, alors que les moyens actuels permettent de viser à l'exhaustivité dans le traitement de l'information contenue dans les documents. Les choix du documentaliste intervenaient dans la détermination des éléments à retenir pour l'indexation ; ils portent aujourd'hui sur la sélection des documents à indexer.

Mais, de l'indexation manuelle à l'indexation automatique, il reste une continuité : la nature du descripteur reste fondamentalement la même, et les structures cognitives de l'esprit humain n'ont pas changé. L'évolution de l'outil rend toutefois nécessaire de fonder la pratique sur une épistémologie explicitée. Le livre de Muriel Amar répond à cette nécessité.

Au centre de la construction, l'indexation est située comme une pratique discursive. Ce ne sont pas les mots qui sont importants, mais les choses que ces mots désignent. Et les mots de la langue ne sont mis en relation avec les choses que par le discours. En outre, le fait de considérer l'indexation comme une pratique discursive est pleinement justifié par le fait que l'univers de discours du documentaliste ne coïncide ni avec ceux des auteurs ni avec ceux des utilisateurs.

Quant au rapprochement de l'indexation et de la vulgarisation scientifique, il ouvre une perspective nouvelle et féconde. Il permet de prendre conscience, sans sombrer dans le désespoir, d'une évidence que la plupart des praticiens et des théoriciens de l'indexation ont refusé de voir : le postulat qui veut que l'univers réel soit le même pour tous les acteurs de la chaîne documentaire, depuis les auteurs des textes sources jusqu'aux utilisateurs, est manifestement faux. Muriel Amar contourne la difficulté par le recours à la théorie des mondes possibles de Kripke, qui permet d'articuler les divers univers de discours.

L'analyse théorique est suivie d'une ouverture sur des propositions pratiques, qui s'écartent des usages habituels. Il n'est jamais confortable de soumettre à un examen

critique la *doxa* officielle de tout un milieu professionnel, même avec un point de vue extérieur, mais l'appartenance de l'analyste à la profession dont il risque de déranger ainsi les habitudes exige un véritable courage intellectuel – pas seulement intellectuel, d'ailleurs. Les compétences de Muriel Amar en épistémologie et en linguistique garantissent la pertinence de son analyse, et son appartenance au corps des conservateurs amplifie la portée de sa démarche. L'efficacité est ici à la mesure du risque accepté.

Michel Le Guern
Professeur à l'Université Lumière Lyon 2

AVANT-PROPOS

L'indexation telle qu'elle a été pratiquée par des générations entières de bibliothécaires et de documentalistes ne serait-elle qu'un simple savoir-faire ?

Certes les entreprises d'indexation, de la plus modeste à la plus aboutie, ont toujours cherché à faire système, à se protéger des subjectivités douteuses, ont toujours redouté la sémantique d'une époque et tenté de trouver, au-delà, les termes aptes à se confronter à l'éternité, à porter en eux, malgré et contre le temps, une forme d'universel et d'immanence. De telles entreprises ont donc toujours été empreintes d'une recherche de rigueur intellectuelle, d'un souci de cohérence, animées par une volonté opiniâtre de constituer des clés d'accès aux savoirs, à la pensée et à l'expression humaines.

Mais, nous dit Muriel Amar dans ce livre fort et exigeant, nous sommes-nous jamais réellement donné les moyens de nos ambitions ? Avons-nous réellement interrogé nos pratiques ? Avons-nous réellement passé nos méthodes au crible de l'analyse scientifique ? Nous qui nous penchons si souvent, justement pour les indexer, sur des travaux scientifiques, avons-vous vraiment cherché à constituer un savoir fondé sur une démarche scientifique ou avons-nous laissé la place à une approximation, à terme fautive ?

Par la seule puissance de son analyse, Muriel Amar donne le sentiment que l'indexation est peu à peu devenue une pratique, on pourrait presque dire une coutume, tellement intégrée dans nos activités fondatrices (mais routinières) qu'elle en devient une évidence – apparente – qui n'est que fort peu interrogée.

Muriel Amar appelle à ce qu'on pourrait appeler une refondation de la légitimité de l'indexation, balayant au passage les arguties de ceux qui, toujours fascinés et dominés par les modes, voudraient faire accroire que la révolution électronique rendrait inutile une telle tâche. Bien au contraire ! Elle ne fait qu'en accroître l'urgence...

Une telle refondation passe par une analyse épistémologique. Muriel Amar défend une approche non instrumentale de l'indexation, fondée sur la linguistique, et clairement différenciée de la recherche documentaire. Le descripteur possède alors une forme d'autonomie par rapport au texte qu'il entend nommer, et doit être apte à *« conjoindre la stabilité de la signification avec l'instabilité de la désignation ».*

Dans un des chapitres les plus séduisants de sa recherche, s'appuyant tout à tour sur Paul Ricœur et Michel Foucault, Muriel Amar cherche les voies d'une indexation conçue comme un aller et retour entre décontextualisation et recontextualisation des documents, passant par une reconnaissance de l'*« épaisseur discursive »* des textes, et capable de construire des descripteurs aptes à refléter tout à la fois la singularité et la pluralité d'un texte, son unicité tout comme son appartenance à tous les textes.

La tentation est grande alors, de répondre à l'invitation de Muriel Amar. Si le descripteur idéal se donne comme ce qui permet *« de circuler dans un espace documentaire conçu a priori comme homogène »,* n'est-il pas alors à lui seul, et d'une manière finalement cohérente, la métaphore même de la bibliothèque ? Ce n'est pas le moindre mérite du livre de Muriel Amar que de remettre l'indexation, dont l'importance est alors pleinement assumée, au cœur des enjeux d'un exercice professionnel.

Martine Poulain
Directrice de Médiadix, Université Paris X

INTRODUCTION

Cette recherche se donne pour objectif de fonder, d'un point de vue théorique, une pratique professionnelle exercée principalement dans les bibliothèques et les centres de documentation : l'indexation, habituellement définie par les praticiens comme « l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document[1] ».

Une étude des fondements théoriques de l'indexation suppose la constitution, à partir d'un objet empirique (la pratique d'indexation), d'un objet scientifique[2].

Cet objectif soulève un certain nombre de problèmes méthodologiques :

(i) concernant l'objet empirique : en quoi les pratiques professionnelles d'indexation peuvent-elles constituer un objet empirique ? Peuvent-elles s'appréhender de façon unifiée ?

(ii) concernant l'objet scientifique : quel point de vue scientifique, quelle science peuvent permettre de construire l'indexation en tant qu'objet scientifique ?

(iii) concernant le type de relation qu'entretiennent objet empirique et objet scientifique : s'il est possible de déterminer des fondements théoriques à l'indexation, qu'est-ce qu'une telle étude peut permettre d'apprendre de l'objet empirique dont elle prétend rendre compte ?

Avant de proposer un cadre de réponse à ces trois questions, nous présenterons succinctement les enjeux d'une étude des fondements théoriques de l'indexation. Nous indiquerons en fin d'introduction le plan suivi dans cette recherche.

---

[1] Norme AFNOR (Association française de normalisation) Z 47-102 (1978), p. 225.

[2] L'opposition proposée ici entre « objet empirique » et « objet scientifique » s'appuie sur l'opposition classique entre *Techné* et *Epistémé* ; les deux notions se distinguent du point de vue de la nature de leur objet : « instable » dans le cas de la *Techné*, « stable » dans le cas de l'*Epistémé*, voir Granger 1993, chapitre II.

I - Enjeux d'une étude des fondements théoriques de l'indexation

Si cette recherche entreprend, en marge des traités théoriques sur l'indexation[1], de repenser la question sous l'angle des fondements théoriques, c'est essentiellement sous « la pression technologique » que connaissent les domaines professionnels de l'information et de la communication. Il s'avère en effet que les descriptions classiques de la pratique d'indexation manuelle dont disposent les professionnels ne leur permettent pas toujours de se situer dans les débats, nouveaux ou moins nouveaux, qui portent par exemple sur l'indexation automatique[2] ou sur l'usage de moteurs d'indexation à l'œuvre sur le réseau Internet[3]. Se multiplie parallèlement une littérature consacrée, elle, à l'automatisation des procédures d'indexation[4], sans que ne soit toujours rendu de façon nette le rapport que l'on peut établir entre l'indexation documentaire, telle que les professionnels la réalisent, et l'indexation automatisée, telle que des systèmes peuvent la produire. D'une certaine façon, on est passé de l'indexation manuelle à l'indexation automatique sans que le niveau d'une appréhension formelle des questions ait été réellement établi. Sur ce point, la problématique de l'indexation ne dépare pas celle des autres objets qui constituent le champ des sciences de l'information et de la communication[5].

Cette carence d'approche théorique de l'indexation n'a pas, semble-t-il, posé de problèmes majeurs jusqu'à ce que :

- d'une part, le nombre croissant de systèmes d'indexation automatisés et la somme totale des coûts engagés n'aient alerté, notamment les pouvoirs publics, sur la nécessité de procéder à une évaluation de ces systèmes ;

- d'autre part, les procédures mises en œuvre sur le réseau Internet imposent de nouveaux modes de traitement des documents qui ne doivent rien à l'indexation telle qu'elle se laisse couramment décrire.

Dans les deux cas, l'absence d'approches formelles de l'indexation constitue un véritable obstacle à l'appréhension comme à la discussion des enjeux mettant en cause l'indexation. Cependant, ce n'est pas toujours la voie d'une entreprise de constitution théorique de l'indexation qui a été retenue ; c'est au contraire une mise à distance de l'indexation elle-même qui semble se dessiner.

A - Place de l'indexation dans les méthodes d'évaluation

L'évaluation de l'indexation a suivi de près la naissance de l'indexation elle-même en tant que pratique professionnelle reconnue. Des outils et méthodes d'évaluation

---

[1] Proposés par exemple par Lancaster [1991] et Fugmann [1993].
[2] Voir, par exemple, sur ce point, Le Moal 1997, p. 380-384. Pendant longtemps, les « logiciels documentaires » (les logiciels dédiés au traitement des documents) sont restés entre les seules mains des informaticiens ; ils passent ensuite dans celles des linguistes-informaticiens avec le développement de l'« ingénierie de l'information ».
[3] Voir, par exemple, sur ce point, Michel 1997, p. 361-363 : « Internet est aussi une vision nouvelle du libre parcours dans l'information. Il pose donc des problèmes à l'industrie de l'information et remet en cause des pratiques développées depuis une vingtaine d'années ».
[4] Pour une synthèse, Sidhom [thèse en cours].
[5] Sur ce point, Le Coadic [1997, p. 516] fait remarquer que « dans le domaine de l'information, la connaissance technique a souvent précédé la connaissance scientifique ».

ont été constitués concomitamment aux outils et méthodes de l'indexation ; les premiers, inspirés des seconds, reposent pour l'essentiel sur une mesure des résultats que l'on peut obtenir à partir de requêtes documentaires[1]. La transposition de ces outils et méthodes à l'évaluation des systèmes d'indexation automatisés s'est avérée problématique[2] et a conduit à la mise en œuvre de programmes de recherche portant sur les méthodes d'évaluation elles-mêmes. Ainsi, pour la période récente, peut-on citer, entre autres, les programmes de recherche suivants :

- sur le plan européen, la Direction Générale XIII (*Language and Research Engineering*) de la CEE a lancé le programme *Eagles*, dont l'un des axes consiste à développer des méthodes d'évaluation pour les produits et services de traitement linguistique de l'information ;

- dans le cadre d'actions multilatérales francophones, l'AUPELF-UREF[3] a engagé une action de recherche concertée (*Amaryllis*) dont le but est de « permettre à la fois à la recherche de progresser et au domaine de se doter d'instruments de mesure rendant possible une comparaison des différentes approches[4] » ;

- en France, une réflexion a été lancée par le CNRS dans le cadre du programme *GRACE*. Parallèlement, le ministère de l'Enseignement supérieur et de la Recherche constitue depuis peu son propre programme de recherche dans le domaine[5].

Les méthodes d'évaluation de l'indexation existantes, sans doute valables lorsqu'elles sont utilisées par les professionnels dans le cadre singulier de leur pratique, ne révèlent pas la même pertinence dès lors qu'il s'agit d'évaluer des systèmes reposant sur des modèles, implicites le plus souvent, de l'indexation : on a bien du mal, dans ces cas, à trouver l'« aune de référence » qui permette de les discuter.

En l'absence d'approche formelle de l'indexation qui permettrait d'évaluer les systèmes d'indexation automatisés sous l'angle d'une évaluation de modèles, la problématique de l'évaluation de l'indexation tend à céder le pas à une évaluation de la capacité des systèmes automatisés à permettre une « bonne » recherche d'information : ce n'est plus l'indexation comme telle qui est évaluée, c'est plutôt un ensemble de procédures diverses censées répondre au même objectif qu'elle[6]. Par conséquent, la question de la « consistance » de l'indexation se pose : l'indexation telle que les praticiens l'exercent n'est-elle qu'une des techniques possibles parmi d'autres ? Dans ce cas, l'indexation peut-elle constituer un objet

---

[1] Sur ce point, voir, dans le glossaire, les entrées « taux de rappel » et « taux de précision ».
[2] Pour une synthèse, on peut se reporter à Sparck-Jones (éd.) 1981. Par exemple, p. 1 : « There is no very good reason to suppose that the conventional methods are best, even in principle, let alone practice » ; p. 3 : « It is arguable that our current understanding of information processing is like of sixteenth century herbolists : it embodies some observation and insight, but lacks detailed analysis and supporting theory ».
[3] AUPELF : Association des universités partiellement ou entièrement de langue française ; UREF : Université des réseaux d'expression française.
[4] AUPELF-UREF 1994, annexe, [p. 1].
[5] Chaudiron 1994, p. 100-104.
[6] Sur ce point, le programme américain TREC (*Text REtrieval Conference*) est exemplaire. On trouvera une présentation des objectifs de ce programme et des résultats auxquels il permet d'aboutir dans Lespinasse 1997.

d'étude ? N'est-ce pas plutôt l'ensemble des procédures utilisées en recherche d'information qui doit alors être analysé ?

B - Place de l'indexation dans le réseau Internet

La constitution du réseau Internet s'est accompagnée de la création d'un ensemble d'outils spécifiques[1], dont certains jouent le rôle de l'indexation documentaire. Ainsi de ceux que l'on appelle les « moteurs de recherche[2] » : ce sont « des bases de données constituées automatiquement grâce à des logiciels appelés robots qui scrutent à intervalles réguliers les serveurs déclarés sur Internet. [...] Ils indexent mot à mot les documents localisés permettant ainsi des interrogations par sujets.[3] » Ces robots qui « indexent » ne procèdent aucunement à une « représentation des concepts » contenus dans les documents, pour reprendre le texte de la norme. Le type d'indexation mis en œuvre ne ressemble en rien à l'indexation que les professionnels pratiquent[4]. Comment interpréter cet état de fait ? L'indexation telle que la pratique professionnelle la définit, la décrit, l'exerce n'est-elle qu'une technique conjoncturelle, liée à un état de la technologie aujourd'hui dépassé, sans être une opération fondamentalement liée au processus de transfert d'information ? L'indexation documentaire est-elle une opération nécessaire ou une technique simplement utile ? Là encore, quel objet le chercheur doit-il retenir pour son étude : l'indexation ? les divers procédés permettant la recherche d'information ?

On aurait tort, nous semble-t-il, d'évacuer trop rapidement les problématiques spécifiques de l'indexation en les diluant dans celles de la recherche documentaire ou dans celles des nouvelles technologies de l'information, c'est-à-dire sans avoir préalablement essayé de formaliser ce qui constitue en propre l'indexation. Alors que « la pression technologique » actuelle tend à laisser les objets se fondre et se confondre (indexation et recherche documentaire, notamment), cette étude entend donner les moyens de « reconnaître » l'indexation sous les aspects différents que peuvent lui donner l'histoire d'une profession comme celle des techniques avec lesquelles elle évolue. Ces moyens, de nature théorique, doivent permettre d'analyser l'indexation en toute généralité, mais aussi de pouvoir capter son évolution, et, pourquoi pas, de la prévoir.

Certes, les pratiques d'indexation telles qu'elles se laissent voir et décrire ne semblent guère sujettes à des généralisations de cet ordre ; il nous semble cependant possible de constituer l'indexation comme un objet scientifique présentant certaines caractéristiques de stabilité.

II - L'indexation, un objet empirique

Qui cherche à étudier l'indexation dispose d'un ensemble d'observatoires de nature différente.

---

[1] Dont le World Wide Web et les techniques associées, entre autres le format HTML (Hyper Text Markup Language).
[2] Comme, par exemple, Alta Vista, Excite ou Lycos.
[3] Lardy 1994, p. 6.
[4] Pour le détail, on peut se reporter à Le Crosnier 1996.

On peut étudier l'indexation sur les lieux professionnels où elle s'exerce et auprès des indexeurs : c'est alors la façon dont les indexeurs indexent qui est analysée. On peut étudier l'indexation telle que les systèmes informatiques la simulent ou cherchent à en simuler les résultats : on s'intéresse alors soit aux procédures techniques (capacité de traitement, temps de réalisation, etc.) soit aux formes de modélisation, implicites et explicites, qui sont à l'œuvre (modélisations de nature mathématique, linguistique, cognitive, etc.). On peut enfin étudier l'indexation telle qu'elle est décrite dans la littérature (normative, didactique, scientifique, etc.).

Une fois déterminés ces différents « lieux » d'inscription de l'indexation (professionnel, technique, discursif), on peut spécifier l'angle d'approche retenu : le processus de l'indexation (les opérations qui la composent), son résultat (souvent appelé descripteur), son objet (le document), ses outils (les langages documentaires), ses supports (on parlera alors de l'indexation de texte, de l'indexation d'image, de l'indexation de carte, de phonogramme, de vidéogramme, etc.).

Le processus lui-même de l'indexation peut être considéré d'au moins deux façons : comme une opération « englobée » ou comme une opération « englobante ». L'indexation peut se concevoir comme une des formes de réalisation, possible parmi d'autres, de l'« analyse de contenu » (à côté de la classification, du résumé et de la synthèse documentaire, etc.). Elle peut également être appréhendée elle-même comme une analyse de contenu, qui se spécifie par le type d'outil qu'elle utilise (classification, langage documentaire, « langage naturel », représentation « conceptuelle », etc.).

Sur la base de cet aperçu, non exhaustif, des aspects de l'indexation, se dégage la diversité des approches possibles. Comme toute pratique sociale, professionnelle, l'indexation ne peut constituer en tant que telle un objet d'analyse. Il est clair qu'une seule approche ne saurait rendre compte de l'ensemble des problématiques de l'indexation. En cela, la pratique de l'indexation constitue typiquement l'objet d'une interdiscipline comme les sciences de l'information et de la communication[1], qui permettent, par le biais de différents types de théorie, de « découper » un aspect du « réel » de l'indexation, et donc de se doter d'un objet empirique observable, sur la base duquel pourra se construire ultimement, par la convergence des approches, un objet scientifique.

III - L'indexation, un objet de quelle science ?

La pratique professionnelle de l'indexation se laisse décrire par le biais de normes, traités, manuels, du point de vue particulier de la pratique elle-même, dans le cadre d'un référentiel proprement documentaire qui maintient l'indexation dans la

[1] Les sciences de l'information et de la communication ont pu être définies comme une « interdiscipline centrée sur l'étude des processus de l'information et de la communication relevant d'actions organisées, finalisées, prenant appui ou non sur des techniques et participant d'actions sociales et culturelles », Comité National d'Évaluation 1993, p. 123. Sur ce point, on peut aussi consulter Têtu [1997, p. 513-516] et Le Coadic [1997, p. 516-523].

complexité de ses manifestations, soumise à une diversité de facteurs de nature hétérogène (institutionnel, technique, historique, etc.).

L'intérêt d'étudier l'indexation dans le cadre des sciences de l'information et de la communication tient au fait que l'ensemble des disciplines[1] par lesquelles elles se constituent comme science permet de disposer d'un ensemble de points de vue théoriques différents et distincts : chacun de ces points de vue propose un référentiel spécifique permettant d'analyser, à un niveau qui lui est propre, l'un des multiples aspects en jeu dans une pratique professionnelle.

Cependant, travailler l'indexation dans le cadre d'une interdiscipline ne présente pas que des avantages[2]. De nombreuses difficultés sont à prendre en considération : comment s'articulent les différents points de vue sur un objet si le « réel » qu'ils permettent de découper n'est pas exactement le même ? Si, du point de vue de la discipline considérée, on peut espérer « tout » voir du phénomène observé, comment savoir ce que ce point de vue permet de faire voir de la globalité de la pratique retenue pour étude ? Pour un objet empirique donné, y a-t-il des approches disciplinaires plus légitimes que d'autres ?

Sur ce point, toute recherche entreprise dans le cadre des sciences de l'information et de la communication doit, nous semble-t-il, contribuer à proposer des réponses à ces questions. Nous essaierons, quant à nous, de prendre en compte cet aspect de la problématique des sciences de l'information et de la communication.

Parmi les différentes disciplines qui constituent les sciences de l'information et de la communication, se trouve la linguistique qui, si elle a été depuis longtemps sollicitée pour l'étude des faits d'indexation, n'a pas toujours été invoquée pour conduire une étude théorique de l'indexation. Comme le note Janik [1985] dans son bilan des rapports entre linguistique et sciences de l'information, la littérature abondante, qui, à la fin des années 60 et aux débuts des années 70, a porté sur les aspects linguistiques des processus documentaires, a surtout privilégié, en fait, le point de vue de l'indexation automatique. Ainsi des ouvrages de Bély et *al.* [1970], Coyaud et *al.* [1972], Cros, Gardin et *al.* [1964], pour les plus connus.

Pour des raisons qui seront développées dans le premier chapitre de cette recherche, nous retiendrons, pour conduire notre étude des fondements théoriques de l'indexation, la linguistique comme discipline de référence. Précisons d'ores et déjà que le choix de cette approche a été déterminé par le travail mené depuis plus de quinze ans par Michel Le Guern et les membres de l'équipe SYDO[3] : les travaux entrepris permettent de disposer d'acquis à partir desquels peuvent se formuler, aujourd'hui, les fondements de l'indexation du point de vue de la théorie linguistique. À bien des égards, cette recherche ne constitue qu'une synthèse,

---

[1] Les sciences de l'information et de la communication se constituent à partir de plusieurs champs disciplinaires, notamment : économie/droit, anthropologie/sociologie, psychologie, linguistique, logique/statistiques/mathématique, histoire/épistémologie/philosophie. Voir par exemple Le Coadic 1994 pour un essai de clarification.

[2] Les sciences de l'information et de la communication constituent un champ de recherche récent (elles existent institutionnellement depuis 1975) dont l'unité épistémologique reste encore en discussion. L'ensemble de ses concepts n'est pas, pour le moment, entièrement établi, Comité National d'Évaluation 1993, p. 87.

[3] L'équipe SYDO (pour SYstèmes DOcumentaires) était composée, à ses débuts, de Alain Berrendonner, Richard Bouché, Sylvie Lainé, Michel Le Guern, Jean-Paul Metzger, Jacques Rouault.

menée du seul point de vue de l'indexation, des études menées par l'équipe SYDO. Nous rappelons donc dans ses grandes lignes le programme de recherche qui a été suivi[1], les acquis qui nous serviront de base de travail et la contribution que voudrait apporter cette étude.

A - Programme de recherche de l'équipe SYDO

Si l'équipe SYDO a travaillé dans le cadre de l'indexation automatique[2], pour laquelle elle a construit un analyseur morpho-syntaxique[3], d'emblée s'est imposée la nécessité de disposer d'un modèle de description formelle de l'unité que l'analyseur visait à extraire des textes[4]. Le descripteur a en effet fait l'objet d'une formalisation détaillée[5], qui s'appuie sur la mise en valeur de ses propriétés spécifiques, référentielle et discursive, justifiant son approche en tant que « syntagme nominal[6] ». Le cadre d'analyse retenu pour rendre compte du fonctionnement particulier du descripteur est double : à la fois linguistique[7] et logique[8]. Dans ce cadre, des études ont pu être menées dont les perspectives d'automatisation reposaient, de façon constante, sur une approche des faits de langue[9].

B - Les acquis

Les études menées dans le cadre logico-sémantique établi par Michel Le Guern permettent aujourd'hui de disposer d'acquis théoriques, qui ne sont pas sans jeter de nouveaux éclairages sur les pratiques d'indexation classiques :

(i) la notion de langage documentaire a pu être mise à distance sur la base d'une étude du rôle des « mots » en indexation : la dimension lexicale propre au

---

[1] La présentation, ici succincte, des travaux réalisés par l'équipe SYDO sera reprise en détail dans la suite de cette recherche.

[2] Le Guern 1994, p. 75 : « Conçu en vue de l'indexation automatique, l'analyseur morpho-syntaxique élaboré par l'équipe SYDO a eu comme premier objectif l'extraction de tous les syntagmes nominaux présents dans le texte à indexer, ces syntagmes nominaux étant amenés à jouer le rôle des descripteurs dans le système d'information. La première tâche a consisté à établir un système de règles qui permette de reconnaître les syntagmes nominaux dans des documents à indexer, le syntagme nominal étant défini, dans une perspective où se croisent la grammaire et la logique, comme la plus petite unité de discours susceptible de servir de base à une relation référentielle autonome ».

[3] La grammaire de l'analyseur a été établie par Berrendonner 1983 et Metzger 1988.

[4] Le Guern 1991a, p. 22 : « Ce dont je suis sûr [...] c'est que vouloir appliquer la linguistique aux traitements automatiques sans se préoccuper de modèles, c'est courir à l'échec, même si le bricolage habile d'un bon informaticien un peu teinté de linguistique peut faire illusion un certain temps. [...] L'informatisation des systèmes documentaires impose la nécessité d'une réflexion théorique sur les opérations qui en constituent les composantes. [...] Le passage de l'indexation manuelle à l'indexation automatique ne modifie pas la nature des descripteurs, mais il oblige à ne plus se contenter d'une approche intuitive et empirique. On peut indexer à la main sans savoir exactement ce qu'est un descripteur ; en revanche, on ne peut pas mettre en place un système d'indexation automatique sans une réflexion préalable sur les descripteurs, et sans une certaine formalisation ».

[5] Le Guern 1984 notamment.

[6] Sur ce point, voir Bouché 1989.

[7] Voir Le Guern 1997, p. 375-379.

[8] Voir Metzger 1997, p. 385-390.

[9] Sans pouvoir être exhaustive, on peut citer, par exemple, sur le traitement des anaphores dans une perspective documentaire, Vidalenc-Sabourin [1989], sur le traitement des conjonctions de coordination, Larouk [1994].

langage documentaire ne correspond pas à la dimension discursive en jeu dans l'indexation ;

(ii) l'indexation se laisse décrire sous la forme d'une extraction d'unités de discours : les notions de « représentation de concepts » et de « traduction de concepts » sont alors à revisiter ;

(iii) la recherche documentaire se laisse, elle aussi, redéfinir[1] : elle a pour finalité non plus l'appariement de « mots », mais plutôt la détermination d'objets particuliers que sont les objets de discours.

Par ailleurs, des rapprochements inédits et fructueux ont pu être opérés entre documentation et terminologie[2], engageant là aussi la recherche dans des voies de nature à spécifier le descripteur sous l'angle des propriétés qu'il partage avec le terme de la terminologie.

Cet important travail de mise au jour des propriétés du descripteur fournit des pistes d'exploration pertinentes pour parcourir le vaste champ des travaux linguistiques à la recherche d'éléments pour fonder la pratique d'indexation du point de vue de la théorie linguistique. En ce sens, les deux dimensions, référentielle et discursive, du descripteur, mises en valeur par Michel Le Guern et les membres de l'équipe SYDO, ont permis de guider notre investigation dans le champ linguistique.

C - Notre contribution

Compte tenu de notre parcours antérieur, nous avons privilégié, dans cette étude des fondements théoriques de l'indexation, le versant linguistique des hypothèses proposées par l'équipe SYDO[3]. Nous avons donc exploré des modes de représentation linguistique de la référence en général et de la référence discursive en particulier, en privilégiant un ensemble de travaux qui s'inscrit de façon plus ou moins lâche dans le programme de recherche proposé par Milner [1989][4].

À partir de ce référentiel linguistique, nous avons repris les hypothèses émises par l'équipe SYDO pour les reformuler dans le cadre des problématiques de l'indexation.

L'essentiel du travail mené sous la direction de Michel Le Guern a porté sur le descripteur vu sous l'angle de la recherche d'information. Nous nous sommes, quant à nous, plus particulièrement attachée au descripteur vu sous l'angle de l'indexation proprement dite, en élargissant la problématique au processus de l'indexation lui-même. Cet angle d'analyse permet de proposer des fondements théoriques concernant en propre l'indexation.

---

[1] Sur ce point, on peut suivre Kuramoto [1995 et thèse en cours].
[2] Le Guern 1989, Mustafa-Elhadi 1989.
[3] Y compris les lectures linguistiques des modèles issus de la logique.
[4] Ce cadre propose une reformulation du programme de recherche proposé par Chomsky. Toutefois, toutes les études linguistiques sur lesquelles nous nous appuierons dans cette recherche ne relèvent ni du même cadre ni du cadre précisément spécifié par Milner. Cependant, elles ne contredisent pas l'option retenue par ce dernier.

IV - Rapport entre objet empirique et objet scientifique

Comme nous le préciserons dans le premier chapitre de cette étude, notre problématique – l'étude des fondements de l'indexation du point de vue d'une théorie linguistique – nous conduit à privilégier, parmi la multiplicité empirique par laquelle peut se capter l'indexation, les discours sur l'indexation[1]. En ce sens, cette recherche porte non pas sur la façon dont les indexeurs ou les systèmes automatisés indexent, mais sur les arrière-plans théoriques sur lesquels reposent de telles pratiques, manuelles ou machinales, d'indexation.

L'étude de cet arrière-plan théorique, tel qu'il se manifeste dans les discours sur l'indexation, permet de constituer, en partie, l'indexation comme objet scientifique : c'est sur la base de reformulation des modèles implicites de la langue que l'on peut spécifier les propriétés linguistiques en jeu dans l'indexation. Sur ce point, nous rejoignons les propositions de Gardin sur le rôle que l'on peut faire tenir aux théories dans l'étude de pratiques non formelles : « À quoi bon prendre la peine de formaliser ou de programmer la collecte et la structuration des données par des voies dont rien ne garantit *a priori* qu'elles se révéleront plus fécondes ou plus « intéressantes », pour l'archéologue ou l'historien, que les voies dites traditionnelles ? N'est-il pas plus raisonnable d'inverser la stratégie, c'est-à-dire de choisir d'abord un certain nombre de théories que la communauté savante ou du moins une partie d'entre elle, tient pour intéressantes ou fécondes, puis d'en donner une version formelle dans l'espoir que l'appareil cognitif ainsi dégagé bénéficiera par construction de ces mêmes qualités, pour d'autres emplois ?[2] »

Notre étude des fondements théoriques de l'indexation présente donc cette particularité d'approcher l'objet empirique « indexation » par le biais d'interrogations issues de problématiques linguistiques : pourquoi est-ce des noms, des unités nominales, qui sont depuis toujours et partout utilisés en indexation, et pas, par exemple, des unités verbales ? Quelle différence y a-t-il entre un descripteur « nom propre » et un descripteur « nom commun », entre un descripteur composé d'un mot et un descripteur composé de plusieurs mots ? Pourquoi les pratiques d'indexation recourent-elles invariablement à la langue, alors même que les discours sur l'indexation ne cessent d'en pointer « l'imperfection », « l'ambiguïté » ? Comment l'indexation appréhende-t-elle la spécificité sémiotique des objets qu'elle manipule, que ce soit les textes qu'elle sélectionne ou les univers de discours qu'elle permet de traverser ?

Cet ensemble de questions, qui ne se pose que dans le cadre d'une approche linguistique des faits d'indexation, permet, par touches, de faire émerger les propriétés de langue sur lesquelles reposent les pratiques professionnelles d'indexation. À ce pouvoir explicatif d'une approche linguistique de l'indexation s'adjoint un pouvoir de nature plus prédictive : on peut déterminer des « manières d'indexer » de diverses natures, dont certaines sont à même de tirer harmonieusement profit des progrès technologiques sans s'y laisser dissoudre.

Le matériau utilisé dans cette recherche des fondements théoriques de l'indexation sera donc principalement constitué d'un ensemble de discours sur la pratique

[1] Cet angle d'étude de l'indexation est couramment retenu ; voir, par exemple, Dubois 1995 ou Van Holland 1995.
[2] Gardin 1991, p. 24.

d'indexation[1], auquel s'ajoute une dimension expérimentale. Nous avons en effet réalisé une enquête auprès de dix organismes documentaires[2] dans le but :

(i) d'analyser le mode d'exploration des sources en indexation : comment l'indexeur construit-il son objet d'indexation, le document ?

(ii) d'étudier le rapport entre le type de document construit et le type de formule d'indexation établi : quelle est l'incidence de la « mise en document » en indexation ?

Les interprétations auxquelles cette enquête peut donner lieu reposent sur les hypothèses que nous permet de formuler une approche linguistique de l'indexation : celles-ci seront spécifiées en cours d'étude ; de même, les conclusions auxquelles on peut aboutir seront ponctuellement rapportées au fil du texte, en fonction des aspects de l'indexation étudiés.

V - Plan de la recherche

Cette recherche procède en trois temps.

- Le premier chapitre est consacré à la formulation de notre problématique. Il propose un cadre qui permet de traiter la question des fondements théoriques de l'indexation. Pour cela, il précise l'objet étudié et la méthode d'analyse retenue, en répondant notamment aux trois questions suivantes :

(i) comment l'indexation peut-elle constituer un objet d'étude spécifique ?

(ii) en quoi une approche en termes de « fondements théoriques » paraît-elle plus adaptée à l'objet étudié qu'une approche en termes de « théorie » proprement dite ?

(iii) pourquoi retenir, parmi l'ensemble des approches possibles, le point de vue de la théorie linguistique ?

- Les deux chapitres suivants sont regroupés dans une première partie intitulée « problèmes théoriques de l'indexation ». Deux problèmes théoriques y sont abordés : la question du lexique en indexation fait l'objet du chapitre II, celle de la référence l'objet du chapitre III. Sur ces deux questions, on examine respectivement le point de vue des professionnels et le point de vue des linguistes, en s'interrogeant sur les zones de distorsion entre descriptions et sur les zones de désaccord entre modes d'appréhension.

- Les chapitres IV et V constituent la seconde et dernière partie de cette étude, intitulée « contribution aux fondements théoriques de l'indexation ». On y propose une reformulation de l'indexation qui repose sur un modèle explicite

[1] Le texte-pivot de cet ensemble de discours est le discours normatif (norme Z 47-100).
[2] Présentée en annexe 1 ; les principaux résultats sont rapportés dans les annexes 2 et 3.

de la langue. Les propriétés linguistiques pertinentes en indexation sont pensées dans le cadre d'un modèle qui permet de les « utiliser » à des fins professionnelles. L'indexation est, dans cette seconde partie, appréhendée sous ses deux aspects de processus et de résultat : le chapitre IV propose de considérer le processus de l'indexation comme un mode d'organisation spécifique des documents, un niveau de « discours » particulier. Le chapitre V reprend la problématique du descripteur sous l'angle d'une approche discursive de l'indexation.

L'articulation de cette recherche est plus précisément présentée à la fin du chapitre I, la formulation de notre problématique permettant de spécifier la logique d'exposition retenue.

Par ailleurs, chacune des deux parties fait l'objet d'une introduction et d'une conclusion spécifiques.

Les termes suivis d'une étoile (*) renvoient au glossaire pages 329 et suivantes.