catalogage : le cœur de métier ne cesse d'évoluer

Analyse du document en tant que support, le catalogage constitue une tâche centrale au sein de la gestion d'un fonds documentaire, y compris numérique. Étroitement cadré par une série de normes et de standards, la démarche de catalogage ne cesse pourtant d'évoluer.

« *base intellectuelle de notre métier de gestionnaire de l'information* » ; ainsi qualifie le catalogage Max Naudi, responsable du pôle expertise bibliothéconomique de la DGES (direction générale de l'enseignement supérieur). « *Je déteste le catalogage, ça m'a toujours rappelé le solfège ou le latin, beurk !* », du côté de Silvère Mercier, conservateur des bibliothèques et blogueur influent (1). Que cela plaise ou non, le catalogage constitue une démarche centrale dans les prérogatives du documentaliste et du bibliothécaire.

Mais qu'est-ce que le catalogage ?

Il s'agit avant tout de l'analyse du document en tant que support. Analyse permettant d'engendrer un catalogue, c'est-à-dire, expliquent Marie-Renée Cazabon et Isabelle Dussert-Carbone (2), « *un instrument efficace permettant d'affirmer si la bibliothèque possède un livre déterminé, défini par son auteur et son titre ; son titre seul si l'auteur n'est pas nommé ; un substitut du titre si l'auteur et le titre ne sont pas appropriés ou insuffisants pour l'identification* ». Concrètement, cataloguer un document, c'est lui associer une notice bibliographique dans un catalogue. Cette notice catalographique se compose de données descriptives en nombre limité – titre, auteur, version, édition, date de publication... – et des accès à cette description, à savoir :

- Le titre propre du document.
- Des *vedettes* correspondant aux portes d'entrée de la recherche, telles les vedettes matières, les vedettes auteurs...
- Une cote, adresse du document sur les rayonnages, ou une URL dans le cas d'une ressource numérique en ligne.

informations pertinentes pour décrire le document

La création d'une notice s'effectue en deux temps : d'abord, la saisie des données descriptives, tâche à faible valeur ajoutée puis, surtout, la description par un ensemble de mots de la teneur du document ; c'est l'indexation. Pour reprendre la terminologie spécifique, la première étape correspond à l'établissement des vedettes auteurs et titres, la seconde, à l'établissement des vedettes matières. « *Déterminer quelles sont les informations pertinentes pour décrire le document,* détaille Max Naudi ; *il s'agit là d'une opération intellectuelle, s'appuyant sur des normes, et structurée par des outils et des langages* ».

La norme répond à la nécessité que connaissent les bibliothèques et les centres de documentation d'échanger les contenus de leurs catalogues. Achat et récupération de notices sont aujourd'hui plus que répandus. Il convient de disposer d'éléments d'évaluation et de normalisation. Né en 1971 à l'initiative de l'Ifla, c'est toute la vocation de l'ISBD (international standard bibliographic description ou description bibliographique internationale normalisée) dont la première édition, ISBD(M), traite des monographies. Elle définit les éléments devant figurer dans une notice, ainsi que le codage permettant d'introduire ces éléments. Elle n'est pas à proprement parler une norme, mais constitue le socle de spécifications à partir duquel chaque pays doit ensuite rédiger ses normes de catalogage. Ainsi, l'ISBD(M) correspond à la norme Z 44-050, et à la Z 44-073 pour sa description allégée. Suivirent rapidement, poussées par le besoin, les publications d'une ISBD(S), pour les publications en série, en 1977, puis d'une ISBD(ER) pour les ressources électroniques – correspondant à la norme Z 44-082.

L'ISBD n'est pas un format informatique, il convient de disposer d'un *langage machine* de catalogage. Le format Marc (machine readable cataloguing) apparaît dans les années 1960. Il permet d'éviter la duplication des efforts, comme recataloguer des fonds déjà catalogués, et favorise l'échange. Il est décliné en formats dérivés – l'Intermarc en France, par exemple, même si l'on converge vers un format universel correspondant à l'Unimarc.

entre normes, standards et protocoles

Seulement voilà, ces formats ne s'appliquent qu'aux documents imprimés, qui sont loin d'être les seules ressources composant le fonds d'un établissement. Pour décrire documents sonores, numériques, vidéos et multimédias, il est nécessaire de faire appel à d'autres standards fondés sur la notion de métadonnées. Il s'agit essentiellement du Dublin Core, schéma

Tournez la page S.V.P.

FORMAT ISBD

Furet, François

Dictionnaire critique de la Révolution française. Institutions et créations [Texte imprimé] / François Furet, Mona Ozouf et collab. – [Nouv.ed.]. [Paris] : Flammarion, 1992 (45-Manchecourt : Imp.Maury). – 349 p. : couv. Ill. en coul. ; 18 cm. – (Dictionnaire critique de la Révolution française : 1992). (Champs : 265).

Notes bibliogr.

Institutions politiques -- France -- 1789-1815 -- Dictionnaires
France -- 1789-1799 (Révolution) -- Dictionnaires

ISBN 2-08-081265-3 (br.) : 42 F

FRBNF35543445

notice bibliographique ISBD lisible par l'usager

FORMAT UNIMARC

000 cam 22 450	Guide (ou lecteur ou en-tête)
001FRBNF355434450000008	Numéro de notice
010 $a2-08-081265-3$bbr.$d42 F	ISBN, reliure, prix
020 aFRb09306652	N° de bibliographie nationale
100 $a19930104d1992 m y0frey50 ba	Données générales de traitement
1010 $afre	Code de langue
102 $aFR	Code de pays
105 $a|||e 00|y|	Données codées (monographies)
106 $ar	Données codées (texte)
2001 $aDictionnaire critique de la Révolution française$iInstitutions et créations$bTexte imprimé$fFrançois Furet, Mona Ozouf et collab	Titre et mentions de responsabilité
210 $a[Paris]$cFlammarion$d1992$e45-Manchecourt$gImpr. Maury	Adresse bibliographique
215 $a349 p.$ccouv. ill. en coul.$d18 cm	Collation
225| $aChamps$v265	Collection
300 $aNotes bibliogr	Note

notice Unimarc en langage machine

de métadonnées générique comportant une quinzaine d'éléments de description formels, et correspondant à la norme Iso 15836. « *Concrètement,* explique Max Naudi, *avec l'utilisation croissante du Dublin Core, on passe d'un catalogage classique avec une notice qui représente le document mais qui lui est extérieur et distinct, à une description à partir de métadonnées, incluses dans le document lui-même* ».

La réalité du catalogage réside largement dans la récupération de notices. Certains logiciels permettent d'interroger plusieurs bases, telles BNF, Decitre et Electre. Est-ce pour autant la fin d'une époque et le passage d'une démarche de médiation, par catalogage classique, à une démarche d'ingénierie documentaire, ainsi que certains le déplorent ? Rien n'est moins sûr.

Protocole de communication informatique client-serveur, la norme Z 3950 permet l'interrogation simultanée de plusieurs catalogues. La notion de catalogue collectif devient par ce biais une réalité, Supposant une grande rigueur – à tout document et quel que soit le nombre d'exemplaires disponibles ne doit correspondre qu'une seule notice – les catalogues collectifs requièrent un référentiel commun. C'est le cas de Rameau (Répertoire d'autorité matière encyclopédique et alphabétique unifié) utilisé notamment par la BNF et par le Sudoc (catalogue collectif des bibliothèques universitaires). Le protocole Z 3950 évolue vers deux autres formats, compatibles avec les technologies web : SRU (search and retrieve via URL) et SRW (search and retrieve via web services).

après rétroconversion et EAD, catalogage 2.0 ?

La numérisation, enfin, concerne aussi les catalogues et impacte fortement l'offre de services des bibliothèques. Elle permet notamment la mise en place d'un catalogue numérique pour l'accès au public : les Opac (open access catalogue). À noter que la transformation des catalogues papier en dossiers numériques et leur intégration soulèvent le substantiel problème de la rétroconversion des données. Dernière perspective en date : le catalo-

⊞ repères

plus de ressources et plus de chiffres

Rigueur et normes sont les deux mamelles du catalogage. Il convient donc de se tenir informé, par des sources qualifiées et officielles. Voici quelques ressources accessibles sur internet.

■ La très institutionnelle *Déclaration des principes internationaux de catalogage* de l'Ifla. Téléchargeable à l'URL suivante :
→ www.ifla.org/VII/s13/icc/imeicc-statement_of_principles-2008_french.pdf
Datée du 10 avril 2008, elle synthétise en une douzaine de pages les conventions et les règles de bonne conduite, mettant en avant le confort de l'utilisateur du catalogue. Un glossaire exhaustif n'est pas le moindre de ses atouts.

■ *Le guide du catalogueur.* Consultable sur :
→ guideducatalogueur.bnf.fr
Il présente les principes de catalogage appliqués par la BNF dans le catalogue BN-Opale Plus. Un ensemble de fiches thématiques illustrées d'exemples donnés à la fois en ISBD, en Intermarc et en Unimarc en fait un vademecum contextuel, précis et précieux.

■ Cours de catalogage : hébergé sur le site de Mediadix de l'université Paris X et réalisé par Jean-Louis Baraggioli, directeur du Centre technique du livre de l'enseignement supérieur, ce cours est consultable à l'URL :
→ netx.u-paris10.fr/eadmediadix/formation/Catalogage/AsiteCatalogage.htm#
Il présente les principales caractéristiques du catalogage ISBD des monographies et des périodiques.

■ Chiffres : la Bibliographie nationale française, indicateur de nombre de notices publiées en France chaque année.
La Bibliographie nationale française rassemble les notices bibliographiques des documents édités ou diffusés en France, et reçus par la BNF au titre du dépôt légal. On y distingue, entre autres, les livres avec 57 294 notices en 2004, puis 62 257 en 2007, les publications en série avec 5 846 en 2007, audiovisuelle avec 24 055 en 2007, la musique et la cartographie.

■ Au 1er septembre 2008, la base du Sudoc présente 8 557 881 notices bibliographiques localisées et 26 242 840 localisations.

un peu d'histoire...

L'histoire du catalogage est avant tout une histoire de tâtonnements et d'expérimentations, une évolution constante vers une rationalisation et une internationalisation de la démarche, aboutissant dans l'après-guerre aux *Principes de Paris.* Ce texte validé en 1961 par la Conférence internationale sur les principes de catalogage sert de base à l'abondante production de règles et de normes édictées depuis. Des catalogues inventorient le corpus des bibliothèques dès l'antiquité grecque, mais c'est avec le catalogue de la Bibliothèque bodléienne – appartenant à l'université d'Oxford – qu'apparaît au XVIIe siècle la première liste ordonnée alphabétiquement permettant de retrouver un ouvrage via plusieurs points d'accès. Après 1791, avec les catalogues confectionnés au dos de cartes à jouer et l'ancêtre d'une première norme avec l'*« Instruction pour procéder à la confection du catalogage de chacune des bibliothèques sur lesquelles les directoires doivent incessamment apposer des scellés »*, ce sont Philippe Otlet et Henri Lafontaine qui, au crépuscule du vingtième siècle, ancrent dans les mentalités la nécessité d'un système international permettant les échanges.

gage en EAD (encoded archival description, en français description archivistique encodée). D'ores et déjà mis en œuvre par l'Abes (Agence bibliographique de l'enseignement supérieur) avec le catalogue et outil de catalogage en ligne Calames [voir *« L'Abes ouvre des perspectives avec Calames »*, *Archimag* n° 213, avril 2008], l'utilisation de l'EAD comme format de catalogage se développe. Offrant tout l'intérêt d'une classification hiérarchique, elle nécessite cependant de réelles compétences en XML.
Et le catalogage 2.0 dans tout ça ? Il n'en est qu'à ses balbutiements, le tagage à la volée, ou folksonomie, constituant plutôt un mode de catégorisation de l'information. Seul élément concret pour l'instant, confie Max Naudi, *« un système d'annotation des notices prévu dans Calames, reste à en fixer les modalités... »*. ■

Guillaume Nuttin

(1) → www.bibliobession.net
(2) Dans *Le catalogage : méthode et pratiques*, tome I. Collection Bibliothèques. Éditions du cercle de la librairie. Paris, 2007.

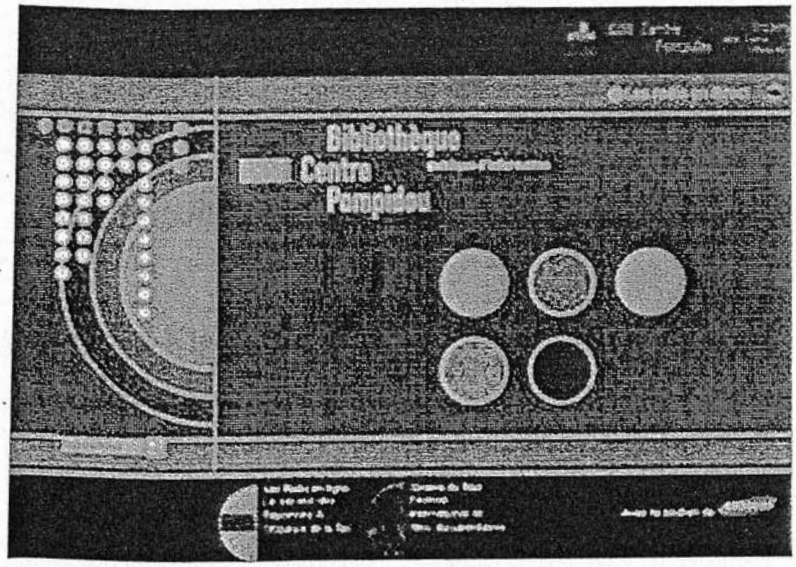
Tournez la page S.V.P.

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/

Alors que la fin – voire la mort – du catalogage est régulièrement proclamée, l'indexation, elle, semble jouir d'une étonnante vitalité[1] : 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 « mot[2] ».

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

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].

Les Dossiers de l'Ingénierie éducative, n° 49, décembre 2004

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 textes[3] en isolant, notamment pour les logiciels d'extraction terminologiques, l'ensemble des unités nominales susceptibles de désigner des objets du monde : ces unités aux propriétés linguistiques spécifiques se nomment des termes[4]. Ce type d'indexation 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 requête[5].

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 utilisateurs de systèmes d'information. C'est pour cette « convivialité[6] » que, dans certains contextes documentaires et pour certains types de besoin d'information, les outils issus de l'ingénierie linguistique sont utilisés[7].

Développer l'intelligence artificielle : l'indexation structurelle

Pour la communauté de l'ingénierie des connaissances, issue de l'intelligence artificielle, la problématique est légèrement différente : l'indexation 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 techniques d'indexation conceptuelle et structurelle aujourd'hui utilisées dans le cadre des travaux du Web sémantique[8], 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'ajouter 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 environnement technique, par exemple[9]). Ces connaissances rendent compte, sous forme d'« annotations », d'une lecture et d'une interprétation expertes des documents (c'est l'indexation conceptuelle) ; ces annotations sont contraintes par des schémas d'annotation, par des « conteneurs de connaissances » (thesaurus, terminologies, ontologies) et par l'appartenance d'un document à un genre (c'est l'indexation structurelle).

« 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. »

Dans ce cadre, l'indexation n'a plus pour seul objectif de permettre 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 documentation 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 communautés de l'ingénierie linguistique et de l'ingénierie 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 utilisateurs 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 spécificités que doivent (ou devraient) prendre en compte les professionnels de l'information.

On retiendra notamment que le passage du support analogique au support numérique permet :

– 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 lecture[10], sous réserve qu'ils relèvent du statut linguistique adéquat (unités nominales référentielles) ;

– de manipuler non plus l'intégralité d'un document mais aussi des segments pouvant, le cas

3. Minimalement, 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 ?

Tournez la page S.V.P.

échéant, être recombinés 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 nom[11].

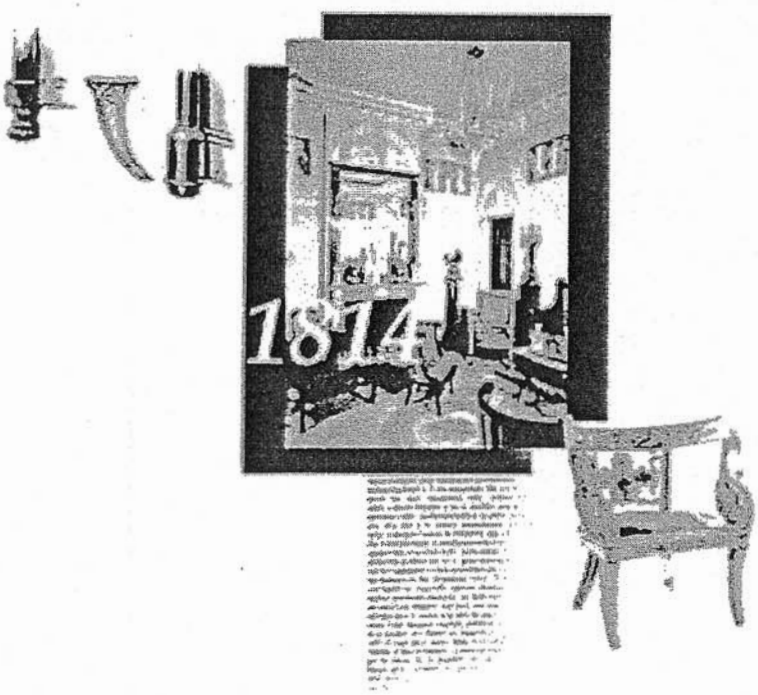
Si ces pratiques d'indexation contrôlée ont fait leur preuve pour le support imprimé[12], que deviennent-elles dans le cas du document numérique ?

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 prédéfinie[13] ne permet pas de donner accès au « texte intégral » du document, c'est-à-dire aux thématiques multiples dont il est porteur[14] ; 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 « objets[15] » que s'ils sont restitués dans le cadre de leur réseau de relations hiérarchiques, associatives ou définitoires[16]. 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 référenciation ?



« ... 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 extra-linguistique, 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 document[17] : 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).

« 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. »

Fondements de la référenciation

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 référenciation). 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 « naturelle » 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.

Du langage documentaire au discours documentaire

Face au double objectif de la thématisation et de la référenciation, 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*. 7e 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 INARecherche*, n° 4, « Interfaces et outils d'analyse et d'indexation ». Séance du 21 juin 1999, compte rendu. www.ina.fr/inatheque/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 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.

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].

Tournez la page S.V.P.

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 document[18].

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 ? ●

[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 oeuvre.
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.

Pourrais-je avoir un nuage de tags sur mon site web 2.0 ? ou faire du neuf avec du vieux

Publié le **17 décembre 2006.**

Petite remise en perspective historique des innovations du Web 2.0. Sous couvert d'innovation, se présentent bien souvent des fonctionnalités ou possibilités déjà existantes, mais un peu rajeunies...

La brumeuse prétendue innovation des nuages de tags

À en croire les commentateurs du Web 2.0, une des grandes innovations mise au service de l'internaute, ce sont les *tags*. Et de fait, on voit fleurir sur les meilleurs sites ou blogs ces "*nuages de tags*" permettant de visualiser la pertinence et le poids des sujets les plus traités sur le site.
Le snobisme consistant à jargonner techniquement en anglais plus que de raison, on ne comprend pas au premier abord que ce fameux système de tags n'est autre que la bonne vieille méthode d'indexation à l'aide de mots-clés, enseignée en bibliothéconomie et en documentation sous le nom d'indexation matière : il s'agit d'apposer des mots-clés décrivant un objet mis à disposition du public : texte, image, son, vidéo...
Une fois de plus, nous découvrons qu'après avoir prophétisé - voire préconisé - la fin des métiers de l'information-documentation dont l'existence serait rendue caduque par l'émergence du Web, celui-ci ne tient que par les techniques documentaires à tous les étages.

Nous passons ici en revue les diverses techniques documentaires utilisées pour sauver, dès l'origine le Web du chaos.

Les balbutiements de l'Internet grand public - c'est-à-dire le Web - furent placés sous le sceau des techniques documentaires classiques, même si d'aucuns ont pu croire que l'ère des techniques documentaires étaient révolues. Illusion d'optique, relayée par quelques penseurs pas forcément au fait des réalités professionnelles (1). C'est ainsi que les tout débuts du Web furent assistés des techniques documentaires réinventées sans le savoir par les pionniers du net.

Les répertoires : classifications inavouées

La première initiative - en termes méthodologiques, car historiquement, les deux naissent en même temps - pour suivre l'extraordinaire démarrage des sites web, fut celle des répertoires. Deux étudiants américains décidèrent de suivre la création de tous les sites Web et de les répertorier. La structure de description des sites ainsi référencés reprenait peu ou prou les éléments essentiels de la description bibliographique : titre du site, auteurs, mots-clés, description...
Mieux : ils allaient structurer la navigation dans le répertoire à partir de "*categories*", comme on dit en anglais, c'est-à-dire tout simplement d'une classification quasi-universelle, à ceci près qu'elle n'avait pas la rigueur des classifications connues en bibliothéconomie. Ainsi s'est développé le répertoire de *Yahoo !*, toujours vivant, même si les dirigeants de la société le masquent pudiquement aujourd'hui, pour des raisons hors de propos ici. Tous les grands répertoires - encore nommés annuaires ou guides. web - obéissent à cette logique classificatoire directement héritée des classifications bibliothéconomiques et documentaires. Comme quoi les métiers de l'infirmation-documentation apportaient quelque chose de bon...

30/04/2009 14:43

Tournez la page S.V.P.

Les moteurs de recherche : logiciels documentaires masqués

L'initiative concurrente a consisté à indexer automatiquement tout le Web mondial, à l'aide d'outils puissants nommés moteurs de recherche (*search engines*). Peu se sont avisés que ces moteurs ne sont en fait que les formes les plus avancées des logiciels documentaires (2), regroupés sous le concept œcuménique d'*informatique de contenu*... De sorte que lorsqu'on cherche une information à partir d'un moteur comme Google, on fait de la documentation sans le savoir...
Bien sûr, le traitement a pris des proportions mondiales et industrielles telles qu'on est loin de l'informatique artisanale des bases de données internes aux centres de documentation. Mais les techniques de base sont exactement les mêmes.
De la sorte, le Web mondial n'est devenu ce chaos très organisé qu'avec l'aide - bien involontaire parfois - des techniques professionnelles issues des métiers de l'information-documentation. L'erreur de bien des professionnels a été de ne pas s'en apercevoir et de croire que l'Internet allait "leur prendre leur travail", plutôt que te tenter de se placer aux positions stratégiques de concepteur de systèmes d'information sur le Web. Et pourtant, il en aurait bien fallu. Rares sont les sites web qui soient correctement décrits et par conséquent, correctement référencés ; parce que non correctement conçus sous l'angle de leur repérage par les moteurs.

Une réalité presque dépassée : les metatags

Il est amusant de se rappeler que les robots des premiers moteurs de recherche exploraient notamment les métadonnées (*metadata*, logées dans des *metatags*) situées dans l'en-tête (partie cachée) d'une page web, dans laquelle le créateur de la page avait la possibilité de fournir ce qu'on nomme une notice bibliographique en bibliothéconomie : Titre, Auteur, Description (petit résumé) et Mots-clés.
Rappelons qu'un groupe de travail s'est penché sur ces questions de métadonnées : le Dublin Core, du nom de la ville des USA (et non la capitale irlandaise) où ils se sont réunis.
Aujourd'hui, si les bons concepteurs de sites pratiquent toujours les métadonnées, celles-ci ne sont plus exploitées par de nombreux moteurs. Pourquoi ? Tout simplement parce que l'indexation qu'elle contient, réalisée par l'auteur du site, n'est pas fiable.
Il y a d'abord les cas de *spoofing* (exagération) : abus de mots-clés intempestifs destinés à piéger les moteurs de recherche et les internautes. Certains ont cru bon d'aligner vingt fois le même mot-clé pour obtenir la meilleure place dans le classement des résultats. Les moteurs ont donc d'abord choisi de ne retenir qu'un maximum de deux occurrences pour chaque mot. Certains ont aussi choisi d'introduire des mots-clés hors sujet, dans le seul but d'attirer de nombreuses connexions (critère pour vendre de la publicité sur son site). Ainsi Pamela Anderson aurait-elle été le sujet de très nombreux sites, si du moins on en croit le champ "mot-clé" des *metatags*...
Il y a ensuite le fait que tout concepteur de site - fût-il bon développeur web - n'a pas forcément les compétences pour bien rendre compte du contenu d'un site par quelques mots-clés soigneusement choisis. On peut oublier certains aspects du site, en exagérer involontairement d'autres. Bref, nous n'allons pas plaider pour le professionnalisme de l'indexation ; s'il existe deux professions pour s'y employer (bibliothécaires et documentalistes), c'est qu'il y a une raison !

De la limite des tags du Web 2.0

Aujourd'hui on nous présente les tags comme une innovation sans précédent, là où on réinvente une technique documentaire séculaire (la classification Dewey est née en 1876). L'innovation réside en effet dans la possibilité de voir s'afficher de manière originale (sous forme de "nuage") les mots-clés les plus usités. Ce système de mots-clés associés aux documents mis en ligne sur les sites personnels ou collaboratifs permettra en effet de mieux les retrouver dans une certaine mesure.
Il n'en demeure pas moins qu'on reste dans un certain amateurisme, puisque les publicateurs de ces informations n'ont pas le savoir-faire pour correctement indexer. Cela prend des proportions tangibles lorsqu'il s'agit d'indexer des photos ou des vidéos. L'indexation de l'image animée est une des choses

les plus délicates. Hormis les données objectives (lieu, date, météo, circonstances objectives telles qu'une fête) - dont les auteurs peuvent déjà omettre certains aspects, l'image véhicule un non-dit que l'analyste d'image sait repérer et décrire, avec une formation appropriée.
Souhaitons en outre fortement que la pratique des tags ne dérive pas en *spoofing* comme les *metatags*, au point qu'on soit obligé de les ignorer. Le bel apport du Web 2.0 perdrait de sa crédibilité, et ce serait dommage.
Une autre difficulté se pose.
La pratique des tags est laissée à l'appréciation de chaque producteur d'information, ce qui est l'esprit même du net. Mais cette liberté même impose des limites à l'efficacité du système.
Tel internaute va produire par exemple un article sur l'informatique documentaire. Il choisira comme tag associé : "informatique documentaire", un autre auteur, à sujet identique, pourrait choisir avec autant de raison "informatisation documentaire", surtout si son article évoque plus la démarche que les outils, cependant qu'un troisième pourrait bien choisir "logiciel documentaire". On touche ainsi du doigt les limites de l'indexation dite libre, en dehors de tout langage documentaire contrôlé (classification ou thésaurus uniformisé) permettant de rattacher une même réalité conceptuelle au même mot, "descripteur" du concept.
Remarquons aussi qu'un même internaute pourra, au fil de ses publications, utiliser un jour "informatique documentaire" et un autre jour "documentation informatisée", tout simplement parce qu'il ne s'est pas souvenu de son ancienne indexation à quelques mois d'écart ; tous les documentalistes ayant pratiqué l'indexation libre connaissent ce phénomène. Une analyse des nuages de tags sur certains sites laisse perplexe à cet égard : des mots-clés très proches coexistent sans qu'on sache ce qui a motivé le distinguo. Leurs auteurs non plus, sans doute... Parfois même, c'est une simple question d'écriture "Web 2.0" et "Web2.0"... ou une simple question de langue : "e-mail" et "courriel".
Cette disparité est en partie rattrapée grâce àl'affichage des mots-clés, sur les fameux nuages, ou lorsqu'il est possible de voir la liste complète des mots-clés utilisés. À condition de se trouver sur le site concerné. Mais pour une recherche via un moteur, ces mêmes mots-clés resteront presque aussi imprécis que le langage naturel contenu dans les textes.

En d'autres termes, la pratique des tags est une innovation qui rend grâce aux techniques documentaires. Même si elle n'est pas aussi affinée, elle permet une certaine amélioration de l'accès à l'information.

Taxonomie et ontologie

Avec l'arsenal des tags, émergent les mots ronflants de taxonomie (ou taxinomie) et d'ontologie. Ces mots savants cachent des réalités là aussi séculaires.
La taxonomie (de *taxis* = placement, mise en ordre et *nomos* = règle, en grec) est l'art du classement d'objets selon une hiérarchie, c'est-à-dire qu'elle est notamment la science des classifications.
Quant à l'ontologie (du grec *ontos* = existant et *logos* = discours), elle permet à l'origine de décrire des objets, intellectuels ou matériels et leurs relations entre eux. C'est donc - pour simplifier, sous notre angle d'étude - la science qui préside à l'élaboration des thésaurus.
Ces termes sont utilisés aujourd'hui par dérivation pour désigner des essais de classification (taxonomie) d'objets, notamment pour s'y retrouver dans les tags. Les ontologies sont utilisées en informatique pour cerner les relations logiques qui existent entre des objets qui doivent être traités par les systèmes automatisés.
Autrement dit, même si nous pouvons paraître réducteur, ces essais mettent au jour les limites de l'indexation libre, dénoncées plus haut. Ils tentent donc *a posteriori* de structurer les notions pour mettre de l'ordre dans les nuages de tags, ou pour améliorer la navigation en introduisant des relations entre les notions.
Somme toute, on réinvente une nouvelle fois les notions de classification et de thésaurus. N'est-ce pas à la fois le plus bel honneur qu'on puisse faire aux pratiques les plus avancées des professionnels de l'information-documentation, mais aussi un constat de malentendu puisque ces métiers devraient - une fois de plus - être présents sur ces terrains sur lesquels on réinvente et remet au jour ce qu'ils pratiquent, même imparfaitement, depuis des lustres ?

Tournez la page S.V.P.

|cc| Didier Frochot — décembre 2006

Notes :

1. Nous pensons notamment à ce grand penseur du management de l'information qui affirmait devant des parterres de managers que les documentalistes ne serviraient plus à rien puisqu'aujourd'hui Internet permettait de tout trouver au bout des doigts... C'était méconnaître à la fois les techniques documentaires et les sites Web en profondeur, comme nous le montrons ici. Repensons aussi à cet article des Échos qui prophétisait, en 1994, la fin des secrétaires, des documentalistes, et même des agents de voyages...
2. Lire : Catherine Leloup, *Moteurs d'indexation et de recherche*. - Eyrolles, 1997.

(cc) Licence Creative Commons. Directeurs de la publication : **Didier Frochot** et **Fabrice Molinaro**.