

La recherche d'information

Panorama des questions et des recherches

Georges VIGNAUX (CNRS-MSH Paris Nord)

- On peut distinguer plusieurs grandes tendances dans la recherche d'information¹ :
- de la dépendance à l'autonomie des usagers,
 - de la maîtrise des stocks à la surabondance des flux,
 - de la validation *a priori* à la validation *a posteriori*,
 - de la rareté et de la distinction à l'explosion et à l'hybridation des outils et des modes de recherche,
 - du modèle de l'accès à celui du traitement de l'information,
 - de la gratuité à la commercialisation de la recherche.

Du côté des usagers : de la dépendance à l'autonomie

C'est sans doute l'évolution la plus significative : depuis les premières recherches des années 60, où l'utilisateur posait sa question au documentaliste qui la transmettait à l'informaticien, jusqu'à l'utilisation actuelle des moteurs de recherche, en passant par l'interrogation des banques de données par le Minitel, les usagers sont passés d'une situation de dépendance totale vis-à-vis des professionnels à une interaction directe avec les outils. Cette autonomisation des utilisateurs est la conséquence directe d'une tendance lourde de l'évolution des outils : la simplification des accès, des interfaces, des procédures. La complexité et l'intelligence technique sont de plus en plus « enfouies » dans la technologie même des outils, et ceux-ci deviennent des « boîtes noires », auto-simplifiantes, utilisables par le grand public (cf le succès de Google). Nous sommes loin d'avoir tiré toutes les leçons de ce phénomène de démocratisation dans l'accès à l'information et de popularisation de pratiques jusqu'alors réservées aux professionnels. Les problèmes de la recherche d'information sont aujourd'hui inséparables des enjeux politiques, culturels, sociaux, liés à l'utilisation des technologies de l'information.

Du côté de l'offre informationnelle

Nous sommes passés de « l'explosion documentaire » des années 60, qui concernait surtout l'information scientifique et technique (essor des banques de données, etc.) à celle du « déluge informationnel » d'Internet. Il s'agit :

- *d'un changement d'échelle*, dans la production documentaire, mesurée désormais en milliards et non plus en millions (sur le Web « visible », *i.e.* indexé par les moteurs de recherche, et impossible à évaluer précisément, le nombre de pages Web serait entre 10 et 15 milliards ; quant au Web « invisible », il serait estimé à 550 milliards de documents !)
- *d'un changement de support*, avec la numérisation généralisée des textes, des sons, des images et de tous types de traces, l'Internet devenant un gigantesque espace « multimédia » ;
- *d'un changement de système éditorial*, le Web étant avant toute chose un vaste système d'auto-publication, permettant à chacun de publier pour le meilleur et pour le pire.

Du côté de la « chaîne de production » de l'information

Contrairement aux centres documentaires protégés et balisés, le Web est un océan ou une poubelle, selon l'appréciation. Ce qui constitue d'ailleurs l'un des enjeux éducatifs les plus forts, c'est bien ce retournement de la validation de l'information : jusqu'alors effectuée « en amont » de la chaîne de production de l'information, d'abord par les chercheurs et les auteurs, qui n'écrivent pas (théoriquement) n'importe quoi, puis par les éditeurs, qui ne publient pas tout

¹ Ce bref panorama prend origine dans la synthèse intéressante établie par Alexandre Serres en 2004 : www.urfist.cict.fr/lettres/lettre34/lettre34-31.html

ce qui s'écrit, ensuite par les libraires, qui ne vendent pas tout ce qui se publie et enfin par les bibliothécaires-documentalistes, qui n'achètent pas tout ce qui se vend, la validation de l'information (*i. e.* l'évaluation, la sélection, le filtrage...) sur le Web : est maintenant généralement reportée sur l'utilisateur, « en aval », avec tous les problèmes et les risques possibles.

Du côté des outils : vers l'hybridation des outils et des modes de recherche

Première observation : nous sommes passés, en deux décennies, d'une relative rareté à une prolifération d'outils de recherche.

Deuxième observation : l'hybridation des modes de recherche et des outils. On peut distinguer, schématiquement, quatre modalités de recherche d'information : la navigation *arborescente* (dans les annuaires thématiques, les classifications), la navigation *hypertextuelle* (dans les sites Web, les encyclopédies), la recherche par *requête sur des mots-clés* dans des champs délimités (l'interrogation des banques de données) et la recherche par *requête sur le contenu* (recherche en texte intégral, moteurs de recherche). A chacune de ces modalités correspondaient des pratiques, des usages de recherche, des outils, jusqu'alors bien distincts. Or l'une des évolutions profondes de la recherche d'informations a consisté à entremêler ces modalités. Depuis quelques années, la mixité entre annuaires et moteurs, combinant recherche arborescente et sur le contenu, et le développement des portails, proposant tous les types de recherche, témoignent de cette imbrication de techniques et de modalités de recherche différentes.

Du côté des processus de recherche

Ces évolutions ont induit une autre transformation profonde, tenant à la fois aux procédures et aux usages de la recherche d'information.

Dans l'univers familier aux documentalistes, c'est-à-dire dans le monde de ce qu'on appelle la « RDI » (Recherche Documentaire Informatisée), les recherches se font avant tout selon la logique booléenne (par l'utilisation des opérateurs booléens, de troncature, éventuellement de proximité) et selon des règles de syntaxe plus ou moins formelles et complexes. La principale caractéristique de la « RDI » tient au fait qu'il s'agit toujours de retrouver des références de documents préalablement saisies : la recherche porte toujours sur un fonds ou une base fermée dont on peut connaître à l'avance le contenu exact ou la composition, et elle fait peu de place au hasard : on sait ce qu'on (re)cherche.

La recherche sur le Web est différente : le contenu est, par définition, impossible à cerner et les modes de recherche sont variés. On peut certes maîtriser toute la gamme des opérateurs, utiliser pleinement les fonctionnalités et les astuces de recherche des outils. Mais quiconque a fait l'expérience d'une recherche sur le Web sait que nombre de découvertes se font souvent par hasard, au gré des navigations de site en site, ou dans la liste des résultats d'un moteur.

Du côté des modèles de la recherche d'information

La question centrale, face au « déluge informationnel », n'est plus tant la recherche elle-même que l'exploitation des résultats. A quoi peuvent servir les milliers de documents trouvés sur Google sur un sujet quelconque ? Comment filtrer le nombre de références, comment exploiter les listes de résultats de manière plus « intelligente », comment obtenir une analyse de tel corpus de données, etc., bref, comment mieux exploiter et gérer les informations : le défi est là¹.

Du côté de l'économie de l'information : de la gratuité à la vente des mots-clés

La nouveauté réside dans cette nouvelle forme d'économie et de marché, apparue autour des outils de recherche privés du Web et des enjeux financiers énormes, à la mesure du trafic généré par ces outils. Liens sponsorisés, liens commerciaux, « *addwords* », etc., les techniques de ce qu'on appelle le « positionnement payant » ne cessent de se développer, ajoutant un nouveau défi pour les usagers : savoir distinguer un lien « sponsorisé » d'un résultat « normal ». Le

positionnement payant consiste en un système compliqué de vente aux enchères de mots-clés, par des sociétés spécialisées (comme *Overture*, *Espotting*) ou certains moteurs de recherche (comme *Google*). Cette vente de mots-clés permettra par exemple à un site commercial, spécialisé dans le voyage, d'apparaître en haut d'une page de résultats pour toute requête comprenant le mot « voyage ». Avec le positionnement payant, c'est la notion même de pertinence qui est atteinte.

Panorama des outils de recherche actuels

Un premier critère, *le mode de recherche proposé*, distinguait autrefois entre les outils par navigation arborescente (comme les annuaires) ou hypertexte (comme les listes de signets), et les outils par requête (comme les moteurs, fondés sur l'utilisation de mots-clés). Cette distinction n'est plus pertinente aujourd'hui, tant l'imbrication est forte sur les mêmes outils.

Un deuxième critère reste toujours valable, en dépit des apparences : celui du *mode d'indexation des ressources*. Selon ce critère, on distingue *les annuaires thématiques*, qui procèdent à un référencement des sites Web (par exemple la partie annuaire de *Yahoo*, *Nomade*, *l'Open Directory*) et *les moteurs de recherche* (*Google*, *Alta Vista*, *Exalead*, *Wisnut*, *YST...*), qui fonctionnent par collecte et indexation automatisées des pages Web (et non des sites). Cette distinction, « historique », est moins nette aujourd'hui, à cause de l'imbrication des annuaires et des moteurs : Google utilise l'annuaire de l'Open Directory, Yahoo a son propre moteur, etc. Mais le critère des modes d'indexation reste essentiel, car il induit des usages et des technologies très différentes. Ainsi un annuaire thématique va-t-il référencer des sites Web, là où un moteur indexera toutes les pages d'un site ; l'annuaire facilitera le défrichage, le premier repérage des ressources dans un domaine ou un secteur défini par l'organisation arborescente proposée, alors qu'un moteur de recherche permettra de trouver un document très précis.

En résumé, la tripartition entre annuaires thématiques, moteurs de recherche et métamoteurs reste une typologie valide. A ces trois catégories d'outils, il faut ajouter deux autres familles : celles des portails et des outils dits annexes. Un portail se distingue notamment des autres outils traditionnels par un ensemble de services personnalisés offerts aux usagers (compte personnel, messagerie, commerce, commande de documents, veille, etc.)ⁱⁱ. Quant aux « outils annexes », il s'agit d'un ensemble d'outils diversifiés, pouvant servir à la recherche d'information et à la veille : « aspirateurs de sites » Web, organisateurs de signets, outils collaboratifs de partage des signets.

Vers la spécialisation généralisée

Un quatrième critère a pris une importance considérable depuis quelques années : *la nature des ressources proposées*. Il s'agit de la distinction classique entre outils généralistes et outils spécialisés.

La spécialisation revêt différentes formes : spécialisation sur un domaine particulier (tourisme, industrie, culture, médecine, sciences exactes, sciences humaines et sociales, etc.)ⁱⁱⁱ, sur une zone linguistique ou géographique, selon la nature des documents (forums, listes de diffusion, bases de données, dépêches d'actualité, bibliothèques électroniques...), selon le type de fichier, selon la nature du média (images, sons)^{iv}.

Les différents niveaux d'analyse linguistique

On peut relever quatre niveaux d'analyse automatisée, correspondant aux quatre premières « couches » d'un texte : morphologique, lexicale, syntaxique, sémantique. A quels niveaux d'indexation se situent les moteurs de recherche ? On sait que lorsqu'on tape un mot-clé sur un moteur, il va chercher dans sa base de données toutes les pages Web contenant ce mot : aucune « intelligence » dans le procédé, mais une simple reconnaissance de chaînes de caractères, qui doivent être identiques. Dans certains cas, le moteur élimine les « mots-vides » (articles, prépositions, etc.). On est dans le domaine de l'analyse morphologique, fondée sur la seule reconnaissance de la forme des mots. Actuellement, la plupart des moteurs fonctionnent encore

à ce premier niveau de l'analyse morphologique (comme Google et Alta Vista).

Quelques moteurs ont poussé l'analyse automatisée jusqu'au niveau du lexique, pratiquant ce qu'on appelle la *lemmatisation* : la réduction d'un mot à sa racine (ou lemme). Du coup, les index sont considérablement allégés, la recherche plus pertinente. La lemmatisation permet également de chercher tous les termes partageant la même racine ou toutes les déclinaisons d'un terme : par exemple, sur *Exalead*, une recherche sur « cheval de course » trouvera non seulement « chevaux de course » mais aussi « course de cheval »^v.

Avec le troisième niveau d'analyse, on passe au stade de la syntaxe, qui permettra de reconnaître des expressions, des groupes nominaux (pollution de l'air, agence de presse, etc.). Assez peu d'outils du Web offrent ces possibilités et on peut citer de nouveau ce moteur français particulièrement innovant, *Exalead*, qui, en plus de la lemmatisation, permet la reconnaissance des groupes nominaux et surtout la proposition de nouveaux mots-clés, par extraction des groupes nominaux du corpus de résultats. La génération automatique de mots-clés constitue d'ailleurs l'une des innovations les plus intéressantes pour l'utilisateur, lui permettant d'affiner ses recherches. On trouve cette fonctionnalité sur quelques moteurs, comme *Alta Vista*, *Teoma*, *Voilà*, à des degrés différents.

Enfin le quatrième niveau d'analyse et d'indexation, celui de la sémantique, concerne la signification d'un texte, par extraction de concepts, de notions. Ce dernier niveau reste peu répandu sur le Web, et se rapproche des pratiques d'indexation avec thésaurus, familières aux documentalistes. L'analyse sémantique est cependant présente sur le Web, selon des méthodes plus statistiques que linguistiques^{vi} : elle concerne surtout le traitement des résultats après une requête et non l'indexation *a priori* des documents. Un exemple intéressant de l'indexation sémantique d'un corpus de textes est fourni par le nouveau service de Google, *News*, dans lequel le moteur propose une « revue de presse » entièrement automatisée, établie à partir des articles et dépêches de journaux.

Les progrès dans les fonctionnalités de recherche et de filtrage de l'information

Ce deuxième domaine d'innovations concerne les interfaces de requêtes. On désigne par là les fonctionnalités, de plus en plus nombreuses, offertes par les outils de recherche (surtout les moteurs)^{vii} pour la gestion des requêtes proprement dites : utilisation des opérateurs booléens et, parfois, de proximité, troncature, équations de recherche avec parenthésage, mais surtout filtrage des requêtes. Certains métamoteurs^{viii} permettent désormais de poser plusieurs filtres sur les requêtes : sur la langue, sur les dates de publication, sur l'espace Internet (Web mondial, francophone...), sur le type de ressources (images, journaux, forums, Weblogs...), mais aussi sur les formats de documents (possibilité de chercher des fichiers PDF, DOC, XLS, PPT...), sur les pages similaires, sur différents champs des pages Web (titre, liens, URL, métadonnées, etc.). La plupart de ces fonctionnalités de recherche restent généralement méconnues des utilisateurs, alors que leur connaissance et leur maîtrise sont l'une des conditions d'une recherche d'information efficace.

Catégorisation, réseaux sémantiques, analyse de contenu

Trois innovations importantes sont apparues depuis deux ou trois ans et concernent la manière dont certains outils de recherche traitent et présentent les résultats d'une requête : *la catégorisation des résultats, les réseaux sémantiques et l'analyse de contenu*.

Mise en œuvre sur le moteur de recherche *Exalead*, et sur le métamoteur *Vivisimo*^{ix}, la *catégorisation dynamique* du résultat des recherches permet de « classer » les documents trouvés dans des catégories, des rubriques porteuses de sens (notamment sur *Exalead*). L'intérêt de cette technologie provient du caractère « dynamique » de cette catégorisation, opérée à partir des caractéristiques réelles du lot de documents trouvés, et non selon des rubriques établies *a priori*. Concrètement, à partir de la requête « cheval de course », *Exalead* a généré, à partir des 68 111 résultats, quatre grandes rubriques (Sport, Commerce et Economie, Régional, Sciences), avec des sous-rubriques (Elevage dans la rubrique Commerce et Economie). (Serres, 2004) Les technologies de catégorisation des résultats réintroduisent ainsi du sens, de la

structuration dans le Web et elles sont appelées, d'une certaine manière, à jouer le même rôle que les thésaurus classiques, avec la différence de taille entre une indexation humaine *a priori* et une indexation automatisée *a posteriori* ...

Deux autres métamoteurs, *Kartoo*^x et *MapStan*^{xi}, ont développé une autre manière de présenter les résultats, non sous forme de rubriques calculées à partir des thèmes propres aux documents, mais sous forme de *cartes, de réseaux sémantiques*, calculés à partir des liens sémantiques entre les pages Web. Au lieu de référer les documents à des catégories thématiques, les pages Web sont reliées les unes aux autres, en fonction des mots-clés qu'elles partagent. Les résultats sont donc présentés graphiquement, sous forme de nœuds et de liens : les nœuds, qui correspondent aux pages Web trouvées, sont de taille variable, selon le degré de pertinence^{xii} des pages Web ; les liens entre les nœuds représentent les relations entre les pages Web, c'est-à-dire leur proximité, leur similarité.

Représentés sous forme de sphères et de liens sur *Kartoo*, de places et de rues sur *MapStan*, ces réseaux sémantiques, parfois difficiles à décoder, offrent plusieurs intérêts pour l'utilisateur : possibilité d'affiner les requêtes (par choix de mots-clés, sur *Kartoo*), de visualiser des liens entre sites Web que l'on n'aurait pas pensé à associer, d'élargir les recherches sur les sites proches, de mettre en évidence (notamment sur *MapStan*) des réseaux d'acteurs sur telle ou telle thématique, avec des indications sur l'importance de tel ou tel site (par le nombre de liens qu'il reçoit)^{xiii}.

Une troisième orientation technologique porte sur *l'analyse automatique du contenu des documents*. Elle est développée notamment par un métamoteur américain, *SurfWax*^{xiv}. Après une requête sur ce métamoteur (qui permet d'interroger près de 500 sources !), une fonction, appelée *SiteSnaps*, offre une sorte de synthèse de l'information sur chaque document trouvé, sous forme de fiche récapitulative : on y trouve ainsi le nombre de mots, de liens, d'images, éventuellement le résumé de l'auteur, les mots-clés de la requête dans leur contexte, les points clés (*Key Points*) de la page. En bref, une sorte d'analyse des documents, permettant à l'utilisateur de mieux faire ses choix, d'affiner et d'élargir sa recherche.

Comme on l'a vu rapidement, ces innovations dans le traitement des résultats induisent des usages différents et offrent des intérêts spécifiques pour la recherche d'information : d'un côté la mise en catégories de documents, de l'autre la représentation cartographique d'un réseau, ou encore l'analyse du contenu.

Vers le « Web sémantique » ?

On ne peut terminer un panorama de la recherche d'information sur Internet sans évoquer ce qui peut représenter une mutation tout à fait majeure, non seulement de la recherche d'information, mais des usages du Web : le « Web sémantique ».

Il s'agit d'un projet de recherche déjà vieux de plusieurs années, lancé par le fondateur du Web lui-même, Tim Berners-Lee, au sein de l'organisation qui préside aux destinées du Web : le W3C (*World Wide Web Consortium*). Le W3C est un consortium créé en 1994, fondé sur trois pôles de recherche internationaux (le MIT, la Keio University au Japon et un regroupement de 18 centres de recherche européens, ERCIM 33), soit au total plus de 500 organisations, universités, entreprises, acteurs importants du Web. Le W3C est donc un acteur essentiel de la « gouvernance » d'Internet, et son rôle est de produire les standards informatiques pour le maintien et l'évolution du World Wide Web^{xv}.

Quels sont les objectifs du Web sémantique ?

Organisation responsable du devenir de la « Toile », le W3C et son président, Berners-Lee, ont été les premiers insatisfaits des nombreux inconvénients du Web, qui ont transformé celui-ci en fourre-tout informationnel. Si le Web originel s'est révélé un fantastique outil pour la production, la publication et la diffusion de l'information, il n'a pu en revanche fournir encore les outils pour structurer et décrire les ressources de manière satisfaisante et permettre un accès pertinent à l'information. Par exemple, les liens hypertextes entre les sites Web, bien que

porteurs de sens pour les humains, n'ont aucune signification utilisable par les machines^{xvi}. On peut citer encore : l'absence ou la faiblesse d'une véritable description des ressources par les métadonnées, la non-exploitation de la sémantique des liens hypertextes par les machines, les limites des outils de recherche, incapables encore d'analyser vraiment les pages Web. Comme l'indique Philippe Laublet (Laublet 2004), le Web actuel est prisonnier d'un paradoxe : « *l'information et les services sur le Web sont aujourd'hui peu exploitables par des machines, mais de moins en moins exploitables sans l'aide des machines.* »

Il s'agit surtout de pouvoir identifier, décrire et indexer les ressources du Web, un peu à l'instar de ce que font les bibliothécaires depuis longtemps à propos des documents.

Sur quelles techniques repose ce projet ?

Le chantier du *Semantic Web* repose sur un empilement complexe de plusieurs « couches » de langages et d'applications informatiques, plus ou moins autonomes. Schématiquement, on peut relever au moins quatre « couches », complémentaires : l'identification, la structuration, la description et la représentation des ressources.

L'identification précise des ressources : les URI

C'est l'objet des URI (*Uniform Resource Identifier*), sorte d'équivalent numérique de l'ISBN pour les livres.

Une structuration logique des ressources : XML

Structuration à la fois homogène et permettant « l'interopérabilité » (mot-clé essentiel du Web sémantique), c'est la « couche » XML (*eXtensible Markup Language*)^{xvii}. Ce « métalangage » (XML n'est pas un simple langage de description et de codage de documents, comme HTML ou PDF, mais une sorte de syntaxe informatique universelle, fondée sur un principe simple : la distinction entre la structure physique d'un document (la mise en page, la typographie, etc.) et sa structure logique (les chapitres, la table des matières), permettant le codage et la description logique de n'importe quel type de ressources (texte, images, données numériques, mathématiques, graphiques).

Une description structurée et pertinente des ressources : les métadonnées

On parle de *métadonnées* à propos de tous les systèmes de description des ressources (depuis les simples balises Meta d'un document HTML jusqu'aux systèmes très élaborés de description, comme le *Dublin Core*^{xviii}, la *TEI*,^{xix} l'*EAD*). Il existe une grande variété de systèmes et de standards de métadonnées et le Web sémantique peut être perçu comme une « surcouche », un cadre général qui vient se superposer à toutes les normes existantes. L'outil développé par le W3C pour le Web sémantique s'appelle le RDF (*Resource Description Framework*) : il s'agit, non d'un nouveau format de métadonnées, mais d'un métalangage, offrant une syntaxe universelle qui permettra aux machines d'échanger des informations de métadonnées incompatibles. RDF distingue trois types d'éléments : un sujet, une propriété, un objet, ou encore une ressource, une propriété, une valeur. Même si ce projet relève encore en partie de la science-fiction, on peut pressentir qu'il changera en profondeur la recherche d'information, en introduisant ce qui manque totalement sur le Web : un système d'indexation portant sur les concepts, les notions.

Une représentation partagée d'un domaine de connaissance : les « ontologies » (OWL)

Une ontologie informatique est une manière de représenter un domaine quelconque de connaissance (disciplinaire, thématique ou autre), sous la forme d'un ensemble de concepts, organisés par des relations structurantes, dont la principale est la relation « est-un » (« *is-a* » pour les anglo-saxons)^{xx}. L'intérêt des ontologies est à rapprocher de celui des thésaurus, avec lesquels elles partagent d'ailleurs beaucoup d'aspects : il s'agit d'outils visant à formaliser un domaine, à permettre à une communauté précise d'acteurs (qu'il s'agisse de bibliothécaires, de

professionnels du tourisme ou de la santé...) de se mettre d'accord sur une représentation commune de leur champ et des concepts qui le constituent, et sur les relations entre les notions. Une ontologie est une « vue sur le monde », ni vraie ni fausse, mais opératoire, partagée et utilisable par les machines. Dans le Web sémantique : les ontologies jouent le même rôle que les classifications, les thésaurus et autres langages documentaires dans les bibliothèques. Ce rôle est essentiel puisqu'il s'agit de permettre aux machines d'établir les liens sémantiques entre différentes ressources.

De nouvelles formes de recherche et d'usage de l'information

Les fondements techniques du Web sémantique ouvrent la voie à de multiples applications nouvelles. Dans la recherche d'information, si les standards RDF et OWL se généralisent sur le Web, de nouveaux moteurs de recherche permettront bientôt de répondre aussi bien à des requêtes génériques, du type : « quelles sont les publications de l'Education nationale consacrées à la documentation ? » qu'à des requêtes beaucoup plus fines, croisant le contenu de plusieurs documents hétérogènes. En bref, le Web sémantique pourrait permettre de surmonter l'hétérogénéité actuelle des ressources du Web, et d'intégrer ces ressources sur une même interface, à partir d'outils simples à utiliser.

Arguments clés

- Les problématiques du traitement de l'information et des nouvelles connaissances numérisées vont s'avérer à terme un enjeu économique, culturel et politique fondamental (cf le projet de très grande bibliothèque numérique de Google).
- Il s'agit en vérité de nouvelles modalités de transformation des modes de pensée dans les modes d'accès à la connaissance.
- L'enjeu pour la recherche française est crucial : il s'agit de résister aux formats imposés, aux catégorisations et aux indexations figées, qui tendent à imposer une « nouvelle » culture mondiale, qui laisse peu de place à l'individu et à la liberté de ses stratégies.

Références

- FOENIX-RIOU, Béatrice. *Recherche et veille sur le Web visible et invisible. Agents intelligents, Annuaire sélectifs, Interfaces des grands serveurs, Portails thématiques*. Paris : Bases, Ed. TEC&DOC, 2001
- LARDY, Jean-Pierre. *Recherche d'information sur Internet. Méthodes et outils*. Paris : ADBS, 2001.
- LAUBLET, Philippe. *Introduction au Web sémantique*. Rennes : URFIST, 2004.
- LEFEVRE, Philippe. *La Recherche d'informations. Du texte intégral au thésaurus*. Paris : Hermès, 2000
- LELOUP, Catherine. *Moteurs d'indexation et de recherche*. Paris : Eyrolles, 1998
- SERRES, Alexandre. *Sélection de ressources sur les outils de recherche*. Rennes : URFIST, 2003. Disponible sur : http://www.uhb.fr/urfist/Supports/ApprofMoteurs_Ressources.htm

Notes et commentaires

ⁱ Citons par exemple le moteur *Exalead*, les métamoteurs *MapStan*, *SurfWax*, *Vivisimo*...

ⁱⁱ Tous les annuaires et moteurs de recherche. (Paris) : disponible sur: <<http://www.enfin.com/>> Répertoire francophone recensant de nombreux annuaires thématiques, généralistes et spécialisés, des moteurs de recherche, des portails, etc.

ⁱⁱⁱ *Internet Search Engine Database*. Cleveland (OH) (USA) : ISEDB.com, 2002-2004. Disponible sur : <<http://www.isedb.com/>> Plus de 1500 outils de recherche référencés, articles, dossiers, actualités.

In-Extenso.org, moteur de recherche en sciences sociales. Voir <<http://www.in-extenso.org/index.html>>

^{iv} *Profusion*, métamoteur spécialisé sur les ressources du Web invisible. Disponible sur : <<http://www.profusion.com>>

^v <http://www.exalead.com/cgi/exalead>. *Exalead* équipe également la plate-forme de recherche d'AOL France : voir : <http://www.aol.fr/>

^{vi} Par méthodes statistiques, on entend notamment le calcul des co-occurrences, c.à.d. le nombre de fois où deux termes apparaissent simultanément dans un texte. Ce type de méthode d'analyse permet d'établir des cartographies des termes et de leurs relations et de dégager ainsi la signification principale, les concepts majeurs d'un texte ou d'un corpus de textes.

^{vii} D'après un travail de comparaison de 7 moteurs de recherche, fait à l'URFIST de Rennes, ce sont *Google*, *AltaVista* et *Voilà*, qui offrent les fonctionnalités de recherche les plus nombreuses : http://www.uhb.fr/urfist/Supports/ApprofMoteurs/ApprofMoteurs_cadre.

^{viii} *Kartoo* ou *Ixquick* : voir également le travail de comparaison de six métamoteurs mené à l'URFIST de Rennes : http://www.uhb.fr/urfist/Supports/ApprofMetamoteurs/ApprofMetamoteurs_cadre.htm

^{ix} Voir <<http://vivisimo.com/>>

^x <<http://kartoo.com>>

^{xi} <<http://search.mapstan.net>>

^{xii} L'indice de pertinence permet de classer les documents selon les mots-clés (nombre, emplacement, « poids » des mots-clés.

^{xiii} Selon cet indice de popularité (le fameux *PageRank* de Google), les pages Web sont classées, non plus selon leur « pertinence » intrinsèque, mais selon leur notoriété sur le Web (cad le nombre et le type de liens pointant vers elles).

^{xiv} <http://www.surfWax.com>. Entre autres fonctionnalités, *SurfWax* propose une fonction linguistique tout à fait originale, le *Focus*, qui permet de préciser les mots-clés d'une requête, en proposant pour un terme les termes synonymes, génériques et spécifiques. Ce *Focus* se présente comme un véritable thésaurus, un outil d'aide à la recherche.

^{xv} C'est le W3C qui a produit et diffusé le standard HTML, le protocole HTTP, le langage XML, et tous les formats et standards propres au Web.

^{xvi} Même si des outils, comme Google ou Teoma, exploitent la structure hypertextuelle du Web, il ne s'agit toujours que de calculs statistiques sur des mots-clés, et non d'une véritable prise en compte de la signification des liens entre sites Web.

^{xvii} Pour un bref historique et une présentation simplifiée de XML, voir : http://www.uhb.fr/urfist/Supports/Rechinfo2/Rechinfo2_cadre.htm

^{xviii} Le *Dublin Core* : système de métadonnées élaboré en 1995 avec la participation de bibliothécaires, permet de décrire une grande variété de ressources sur internet, à partir d'un ensemble de 15 rubriques de description.

^{xix} La TEI (*Text Encoding Initiative*) permet l'échange de données textuelles, mais aussi d'images et de sons, et vient des communautés scientifiques, notamment d'informatique et de linguistique.

^{xx} L'ontologie, dans son acception philosophique habituelle, signifie la « science de l'être », portant sur les concepts généraux, tels que la substance, l'existence, l'essence, ou encore « la partie de la métaphysique qui étudie les êtres tels qu'ils sont en eux-mêmes, et relativement à leur cause » (d'après

Nouveau vocabulaire des études philosophiques, S. Auroux et Y. Weil, Hachette, 1975).