# Recherche d'information sur Internet : où en sommes-nous, où allons-nous ?

Alexandre Serres

**Alexandre SERRES**
**Maître de conférences en Sciences de l'Information et de la Communication**
**Co-responsable de l'URFIST de Rennes**

# Recherche d'information sur Internet : où en sommes-nous, où allons-nous ?

Tout le monde sait que l'accès et l'utilisation d'Internet dans les CDI posent des problèmes difficiles et des questions souvent cruciales aux documentalistes : à quelles conditions doit-on permettre l'utilisation d'Internet dans un collège, faut-il réglementer, comment former les élèves à mieux interroger les moteurs de recherche, comment leur apprendre à identifier et à évaluer l'information, etc. L'une des réponses à ces nombreux défis réside dans la formation professionnelle des documentalistes, dans leur maîtrise des techniques et des outils, devant lesquels ils se sentent parfois désarmés, voire dépassés par les élèves de la « génération Google-Napster », qui « surfent » en toute légèreté sur le web. Car si les documentalistes ne sont pas (et n'ont pas à être) des spécialistes de la recherche d'information mais des usagers avertis et des formateurs compétents, ils ne peuvent se désintéresser des évolutions techniques qui touchent les outils et conditionnent, en large part, les usages des élèves. Cet article ne vise donc rien d'autre que de tenter de fournir quelques éléments de compréhension de ces évolutions qui donnent le tournis.

## Le difficile état des lieux

Explosion des outils, incessants progrès des techniques de recherche, nouvelles applications (comme les *weblogs*[1], le *RSS*[2]…), travaux du Web sémantique… : l'univers de la recherche d'information sur Internet est en constante ébullition, en transformation permanente depuis 10 ans et il est très difficile, voire impossible, de faire un arrêt sur images ou de dresser un tableau complet, un état des lieux clair et ordonné des outils, des technologies et des évolutions qui les concernent.
Cette difficulté de l'observation est sans doute l'un des nouveaux traits du paysage de la « RII »[3], d'autant plus évident pour les « anciens » de la documentation, qui se souviendront des premières banques de données des années 80, du Minitel et des logiciels documentaires primitifs, composant alors un monde documentaire limité et maîtrisable par les documentalistes.

---

[1] Littéralement, un *weblog* est un journal de bord sur le web. On parle aussi de "blog", de "joueb" (pour journal web). Un weblog est un journal personnel, publié sur le web, comportant des commentaires et des listes de liens, régulièrement mis à jour, et permettant l'interactivité. Il y aurait plus de 2 millions de weblogs, de toutes catégories : journal intime, carnet de bord, tribune libre politique, forum communautaire, etc.
Voir *L'ABC du blog* : <http://www.pointblog.com/>

[2] RSS : *« Rich Site Summary »*, ou « *Really Simple Syndication* ». Le RSS est un format informatique permettant la "syndication », ou l'agrégation, de contenu", c.a.d. la possibilité de publier automatiquement sur un site web des informations issues d'un autre site : dernières nouvelles, articles, nouveautés...

[3] Pour des raisons de commodité, je parlerai de RI pour Recherche d'information et RII pour la Recherche d'information sur Internet

Serres, Alexandre. Recherche d'information sur Internet : où en sommes-nous, où allons-nous ? [en ligne]. *SavoirsCDI*, Juin 2004. 14 p. Disponible sur : https://www.reseau-canope.fr/savoirscdi/index.php?id=1052

Aujourd'hui, quel professionnel de l'information peut se prévaloir d'une connaissance complète de la RI ?
Notre propos n'est donc pas l'exhaustivité du panorama, ni l'approfondissement de la compréhension des techniques en jeu, mais plutôt d'essayer de discerner quelques unes des grandes tendances de la recherche d'information, en cours et à venir, et d'apporter quelques pistes de réflexion sur les conséquences de ces évolutions pour les documentalistes des CDI.

## Sept grandes tendances de la recherche d'information

Avant de pointer ces évolutions en cours et à venir, il n'est sans doute pas inutile de revenir brièvement en arrière et de mesurer le chemin parcouru ; le regard rétrospectif, le souci de l'histoire, toujours nécessaire, sont de plus en plus oubliés ou minorés aujourd'hui, notamment à cause de l'emballement technologique, où une innovation chasse l'autre tous les six mois, en effaçant les traces des techniques passées. Sans faire ici un historique des outils et des techniques, nous évoquerons quelques unes des grandes tendances qui peuvent résumer les principaux bouleversements de la recherche d'information depuis les débuts de l'informatisation documentaire, il y a plus d'une quarantaine d'années.
En prenant en compte les différents « composants » de la recherche d'information, nous en avons relevé sept, sans aucune prétention d'exhaustivité, que l'on peut résumer ainsi :

- 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 « retrouvage » booléen à la « sérendipidité »
- 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 totale à l'autonomie relative**

Il s'agit là sans aucun doute de l'évolution majeure, profonde, de la recherche d'information, qui résume toutes les autres : depuis les premières recherches « en différé » 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 croissante des utilisateurs est la conséquence directe d'une tendance lourde de l'évolution des outils : la simplification d'usage. Simplification des accès, des interfaces, des procédures... : toute la complexité et l'intelligence technique sont de plus en plus "enfouies" en amont, 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 irréversible, qui marque par ailleurs un extraordinaire phénomène de démocratisation dans l'accès à l'information et de popularisation de notions et de pratiques jusqu'alors réservées aux professionnels. Les problèmes de la recherche d'information concernent aujourd'hui tout le monde et sont inséparables des questions et enjeux politiques, culturels, sociaux…, liés à l'utilisation des technologies de l'information, à la « fracture numérique ». L'usage de masse des outils de recherche pose également de nombreux (et relativement nouveaux) problèmes, sur les modifications des pratiques informationnelles

(comme le « zapping »), sur les nouveaux modes de connaissance induits par les logiques des outils de recherche[4], sur les dangers de « l'info-pollution », etc.
Mais si les usagers, et donc les élèves, deviennent de plus en plus autonomes, faut-il pour autant les laisser « seuls face à Internet » ? Quelle est la nature de cette autonomie ? Nous reviendrons sur cette question de l'autonomisation (surtout procédurale) des usagers, qui pose la question de leur formation intellectuelle.

- **Du côté de l'offre informationnelle : de la maîtrise des stocks à la surabondance des flux :**

Du côté de l'offre informationnelle, nous sommes passés de la métaphore de « l'explosion documentaire » des années 60, qui concernait surtout l'information scientifique et technique (essor des banques de données, accroissement des revues scientifiques, etc.) à celle du « déluge informationnel » d'Internet. Mais le changement est loin d'être métaphorique, car il s'agit tout à la fois :

- **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 ce qu'il veut, pour le meilleur et pour le pire.

Tous ces aspects sont connus et il est inutile d'y revenir. Mais ce « déluge informationnel » a entraîné une conséquence, tout à fait essentielle pour les documentalistes, qu'il ne faut jamais perdre de vue : le renversement de la problématique de la validation documentaire.

- **Du côté de la « chaîne de production » de l'information : de la validation *a priori* à la validation *a posteriori* :**

On sait que contrairement aux bibliothèques et aux CDI, ces espaces documentaires protégés, surveillés, balisés et aux allées (généralement !) bien droites, le web est une jungle, un océan, un fouillis ou une poubelle, selon l'appréciation. Ce qui est patent, et 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 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 (terme générique sous lequel on mettra l'évaluation, la sélection, le filtrage…) s'opérait à différents niveaux, par différents acteurs, selon différentes modalités et pour différentes finalités. Ce schéma, toujours valable dans le monde « traditionnel » de l'édition ou de la production scientifique, n'est plus celui du web : la validation de l'information est ici généralement reportée sur l'utilisateur, « en aval », avec tous les problèmes, les risques et les dégâts possibles. Le web ou la crise des médiateurs…

- **Du côté des outils : de la rareté à l'explosion, de la distinction à l'hybridation des outils et des modes de recherche**

[4] Voir le récent débat virtuel sur ces questions, organisé par la BPI : <http://debatvirtuel.bpi.fr>

Une première observation très simple frappe l'observateur, devant le paysage actuel des outils et des méthodes de recherche : nous sommes passés, en deux décennies à peine, d'une relative pauvreté, ou rareté des outils, à une abondance, **une explosion, une prolifération d'outils de recherche** ; dans les années 80, par exemple, la recherche d'information se limitait aux logiciels documentaires (environ une vingtaine), aux logiciels des serveurs de banques de données, et à quelques outils documentaires spécifiques. Aujourd'hui, les outils de recherche, uniquement sur Internet, se comptent par milliers, il en meurt et il en naît chaque année plusieurs dizaines et la diversité, dans la spécialisation, est infinie

Deuxième observation : **l'imbrication, 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 CD-ROM, les encyclopédies), la **recherche par requête sur des mots-clés dans des champs délimités** (l'interrogation des banques de données, des catalogues…) 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, mais aussi des outils, jusqu'alors bien distincts. Ainsi, aux débuts du web, la distinction entre annuaires et moteurs était-elle parfaitement claire, du seul point de vue de la modalité de recherche. Or l'une des évolutions profondes de la « RI » a consisté à entremêler ces diverses modalités ainsi que les outils sur lesquels elles s'appuient. Depuis quelques années, la mixité entre annuaires et moteurs, combinant recherche arborescente et sur le contenu, ou le développement des portails, proposant tous les types de recherche, témoignent de cette imbrication d'outils, de techniques et de modalités de recherche différents, ajoutant parfois de la confusion au paysage et rendant les distinctions de plus en plus difficiles. Pour autant, les typologies d'outils restent fondamentales à maîtriser et à expliquer aux élèves, pour sortir des « apparences de l'écran ».

- **Du côté des processus de recherche : du « retrouvage » booléen à la « sérendipidité »**

Toutes ces évolutions, touchant aussi bien à l'offre, aux outils et aux modalités de recherche sur Internet, ont induit une autre transformation profonde, tenant à la fois aux procédures, aux processus mêmes et aux usages de la recherche d'information.

Dans l'univers documentaire, familier aux documentalistes, des bases de données, des catalogues de centres documentaires ou de bibliothèques, c.a.d. dans le monde de ce qu'on appelait 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 (cf les équations de recherche). Mais 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 (quelle que soit la taille de cette base), dont on peut connaître à l'avance le contenu exact ou la composition, et elle fait peu de place au hasard et à l'intuition. On sait ce qu'on (re)cherche. Et les notions de bruit et de silence s'appliquent parfaitement à cet univers documentaire.

La recherche sur le web offre à cet égard un tout autre aspect : le contenu est, par définition, infini, impossible à caractériser, à cerner et les modes de recherche sont variés, comme on l'a vu (logique booléenne, mais aussi navigation hypertexte, recherche sur le texte intégral…).

On peut certes (et on doit !) 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 les plus belles découvertes se font souvent par hasard, au gré des navigations de site en site, ou dans la liste des résultats d'un moteur. On ne sait pas ce qu'on cherche.

Un terme, très employé aujourd'hui sur le web, désigne ce nouveau type de recherche, la **sérendipité**, traduction (non encore officielle, semble-t-il) du terme d'origine, *serendipity*, définie comme « *la découverte par chance ou sagacité de résultats que l'on ne cherchait pas* » (sur la sérendipidité, voir les articles de Ertzscheid-Gallézot et S. Catellin[5])
Fondée sur l'intuition, l'association d'idées, mais aussi l'abduction, la sérendipidité, qui caractérise également le mécanisme de l'invention et de la découverte scientifique, est devenue l'un des modèles de la recherche d'information sur internet, remettant au premier plan les questions de la navigation hypermédia, de l'apprentissage par exploration, de la conversion des informations en connaissances… Il ne faut pas y voir pour autant la fin de la nécessaire maîtrise des principes et des procédures de recherche ou l'inutilité de tout savoir technique sur les outils de recherche.

- **Du côté des modèles de la recherche d'information : de l'accès au traitement de l'information**

Selon une étude californienne, 16 % du temps de travail en entreprise serait consacré à rechercher l'information. « *Soit, si l'on transpose à un travailleur français soumis aux 35 heures, un peu plus de 5 heures 30 par semaine…* »[6] Qui peut encore nier que la recherche d'information n'a pas un coût économique et ne représente pas un enjeu majeur ?
Et la question centrale aujourd'hui, face au « déluge informationnel », n'est plus tant la recherche elle-même que le traitement, l'exploitation des résultats. A quoi peuvent servir les 3OO OOO documents trouvés sur Google sur un sujet quelconque ? Comment filtrer, comment réduire le nombre de références, comment exploiter les listes de résultats de manière plus « intelligente », comment avoir un aperçu du contenu de telle page web, comment obtenir une analyse de tel corpus de données, etc., bref, comment mieux exploiter et gérer les informations : le véritable défi est là. Un certain nombre d'outils de recherche, parmi les plus innovants[7], travaillent depuis quelques années sur ces problématiques et ont apporté des innovations parfois spectaculaires (cf plus loin). Et il est intéressant d'observer une sorte de chassé-croisé entre les deux univers technologiques, que sont les logiciels documentaires et les outils du web : d'un côté les logiciels de gestion documentaire « en local » (de type Superdoc, BCDI, JLB, etc.) évoluent de plus en plus vers les technologies du web, basculant parfois complètement sur une interface web et rendant transparentes la création et la diffusion de bases de données locales sur le web ; de l'autre côté, certains outils de recherche du web ont intégré des fonctionnalités de gestion documentaire, notamment d'analyse de corpus, ou des notions issues de la documentation (comme les thésaurus) et les appliquent à l'univers du web

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

---

[5] ERTZSCHEID, Olivier, GALLEZOT, Gabriel. *Chercher faux et trouver juste, Sérendipité et recherche d'information.* Congrès de la SFSIC, Bucarest, Juillet 2003. Disponible sur : http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/06/89/sic_00000689_02/sic_00000689.html
- CATELLIN, Sylvie. Sérendipité, abduction et recherche sur Internet. In *Emergences et continuité dans les recherches en information et communication*, Actes du XIIe Congrès national des SIC, UNESCO (Paris), 10-13 janvier 2001. Paris : SFSIC, 2001

[6] REMIZE, Michel. Recherche et gestion de l'information : convergence vers le métier documentaire. *Archimag*, n° 172, mars 2004, p. 44.

[7] Citons par exemple le moteur *Exalead*, les métamoteurs *MapStan*, *SurfWax*, *Vivisimo*…

Enfin, une autre tendance lourde, parfois insuffisamment connue, se développe depuis plusieurs années et est en passe de transformer en profondeur le paysage de la recherche d'information : l'inscription, au cœur même des procédures de recherche, de la commercialisation. Il ne s'agit pas de la fin d'un prétendu âge d'or de la gratuité de l'information, qui n'a jamais vraiment existé. L'information a toujours eu un coût (le plus souvent supporté par le fournisseur) et souvent un prix pour l'utilisateur. Les banques de données professionnelles et scientifiques sont le plus souvent payantes (et souvent très chères), et ce depuis leur naissance.
La nouveauté réside ici 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* »…, les techniques de ce qu'on appelle le « *positionnement payant* » ne cessent de se diversifier et de se développer, ajoutant un nouveau défi et un nouveau risque pour les usagers parfois distraits : savoir distinguer du premier coup d'œil un lien « sponsorisé » d'un résultat « normal ». Rappelons que 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.
Après ce rapide recensement de quelques unes des tendances lourdes de la recherche d'information, quel est, aujourd'hui, le paysage plus spécifique des outils de recherche et quelles sont leurs principales évolutions ?

## Panorama des outils de recherche actuels

Le temps des typologies des outils, simples et visibles, paraît loin et il devient difficile de s'y retrouver dans un paysage sans cesse mouvant : mixité des annuaires et des moteurs (cf par exemple le dernier exemple en date du lancement du moteur de Yahoo, YST (pour *Yahoo Search Technology*), floraison d'outils spécialisés, rachats et disparitions d'outils (Alta Vista racheté par Yahoo, fermeture de LookSmart), diversification des technologies, le tout dans un marché dominé par l'écrasante suprématie de Google (qui réalise tout seul 56 % du trafic mondial généré par les moteurs de recherche[8] !)…

### Plusieurs typologies possibles

Pour une compréhension théorique des outils de recherche, le recours aux typologies reste le premier pas, pour tenter de mettre un semblant d'ordre dans un paysage confus.
Plusieurs critères, différents et complémentaires, permettent de répartir les outils de recherche. Eliminons un premier critère : **le mode de recherche proposé**, qui distinguait 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 très pertinente aujourd'hui, tant l'imbrication des modalités de recherche 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

[8] D'après la *Lettre "Actu Moteurs"*, n° 288. Mai 2004.

procèdent à un référencement et une description humaines des sites web (par exemple la partie annuaire de *Yahoo, Nomade, l'Open Directory*…) et **les moteurs de recherche** (*Google, Alta Vista, Exalead, Wisenut, YST*…), qui fonctionnent par collecte et indexation automatisées des pages web (et non des sites). Cette distinction, « historique » car elle a longtemps structuré le monde des outils, est moins nette aujourd'hui, à cause de **la mixité, 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 ressources, 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. Autrement dit, les deux familles se prêtent à des utilisations complémentaires : pour connaître la liste des journaux présents sur le web, la navigation dans un annuaire sera recommandée, alors que vous y trouverez difficilement un support pédagogique sous Power Point, en français, paru en 2002 et traitant du fonctionnement des ordinateurs…

Le critère **du fonctionnement interne des outils** reste également pertinent, et permet la distinction entre les différentes familles d'outils, non seulement celles des annuaires et moteurs, mais aussi des **métamoteurs, des portails, des outils de veille, des outils annexes**. D'un côté, des outils « de première main » pourrait-on dire, qui possèdent leur propre base de données de ressources (qu'elles soient collectées « humainement » ou automatiquement) et leur propre module d'interrogation, de l'autre, des outils « de seconde main », qui n'ont pas de base de données en propre, mais seulement un module d'interrogation, qui exploite les bases des moteurs et des annuaires. La famille des métamoteurs de recherche, apparue presque en même temps que les deux premières catégories, reste un ensemble d'outils très riche, foisonnant, où les innovations techniques sont nombreuses et spectaculaires. Nous y reviendrons. On y associe généralement une autre catégorie d'outils, les « agents intelligents ». Ces outils sophistiqués, proches des métamoteurs, servent à des utilisations très précises : analyse, veille, comparaison de prix… Un agent intelligent peut se définir comme un outil capable d'autonomie, de collaboration (avec d'autres outils) et d'adaptation à son environnement. Outre certains métamoteurs grand public (comme Copernic, proche des agents intelligents), ces outils restent surtout utilisés « en interne », par les entreprises ou les universités, et sont capables de filtrer, d'analyser, de cartographier, de résumer… des corpus d'informations hétérogènes.

En bref, la **tripartition de base entre annuaires thématiques, moteurs de recherche et métamoteurs** reste une typologie structurante et valide, essentielle à comprendre et à faire comprendre aux élèves.

Mais à ces trois catégories d'outils aisément reconnaissables, il faut ajouter deux autres familles, moins faciles à définir : celles des portails et des outils dits annexes. Qu'est-ce qu'un « portail » ? Sylvie Dalbin le définit comme une *« ressource accessible via Internet, constituant un point d'accès unique, simplifié, facile d'emploi et unifié, pour un public cible, à des ressources (services, produits) électroniques distantes, variées et hétérogènes* »[9]. Sans approfondir ici cette notion, disons qu'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.).

Quant aux « outils annexes », on range sous cette appellation vague un ensemble d'outils diversifiés, pouvant servir à la recherche d'information et à la veille de manière annexe : « aspirateurs de sites » web, organisateurs de signets, outils collaboratifs de partage des signets…

[9] Sylvie Dalbin, Instruments de recherche sur le Web, in *La Recherche d'information sur les réseaux, cours INRIA 2002.* ADBS, 2002, p. 23

On le voit, la distinction entre les familles d'outils est souvent problématique, Yahoo symbolisant à lui seul cette hybridation des outils, puisqu'il est à la fois un annuaire thématique (le premier annuaire mondial), un portail (le premier à avoir introduit la « portalisation ») et désormais un moteur de recherche à part entière…

**Vers la spécialisation… généralisée**

Un **quatrième critère** transcende toutes ces catégories et a pris une importance considérable depuis quelques années : **la nature des ressources proposées**. Il s'agit là de la distinction, classique en documentation, entre **outils généralistes et outils spécialisés**, quel que soit le type de spécialisation.
Trois remarques à propos de ce critère :

- il est devenu majeur aujourd'hui, avec l'extraordinaire développement des outils spécialisés. On trouve désormais sur Internet des outils de recherche pour à peu près tous les supports, tous les thèmes, toutes les régions, toutes les applications possibles et la **spécialisation est bel et bien une autre tendance « lourde » de la recherche d'information**. Ce qui pose le problème de la recherche de ces outils spécialisés eux-mêmes, justifiant ces « méta-outils », ou répertoires d'outils, comme *Enfin.com*[10] (qui recense 4893 outils, dont 2970 outils de recherche francophones) ou *Internet Search Engine Database.*[11]

- La spécialisation revêt **toutes les formes possibles** : **spécialisation sur un domaine particulier** (tourisme, industrie, culture, médecine, sciences exactes[12], sciences humaines et sociales[13], etc.), sur **une zone linguistique ou géographique**[14]**,** selon la **nature des documents** (forums, listes de diffusion[15], bases de données, thèses, dépêches d'actualité, bibliothèques électroniques…), selon **le type de fichier**, selon la **nature du média** (images, sons…)

---

[10] ***ENFIN.*** *Tous les annuaires et moteurs de recherche*. (Paris) : IDF.net, 1999-2003. Disponible sur WWW : <http://www.enfin.com/> Répertoire francophone recensant de nombreux annuaires thématiques, généralistes et spécialisés, des moteurs de recherche, des portails, etc. Accès par requête et par navigation dans cet annuaire d'outils

[11] ***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. Accès par moteur et annuaire.

[12] Cf par exemple ***Scirus***, moteur de recherche spécialisé en sciences exactes : Scirus - for scientific information Disponible sur : <http://www.scirus.com/>

[13] Par exemple In-Extenso.org, moteur de recherche en Sciences sociales. Voir <http://www.in-extenso.org/index.html>

[14] Voir par exemple Breizhoo.fr, le moteur de recherche spécialisé sur la Bretagne. Disponible sur : <http://www.breizhoo.fr/>

[15] Voir Francopholistes, l'annuaire des listes de diffusion : <http://www.francopholistes.com/>

- Elle **touche toutes les familles d'outils** : annuaires, moteurs de recherche, métamoteurs[16], portails… Chaque catégorie d'outils se partage entre outils généralistes et spécialisés.

## Trois domaines d'évolution technique dans les outils de recherche

Au risque d'une présentation simplifiée des évolutions techniques des outils, il nous semble judicieux de privilégier trois domaines :

- **l'intégration des techniques du TAL** (Traitement Automatique des Langues)
- **les progrès dans les fonctionnalités de recherche et de filtrage de l'information**
- **la diversification des méthodes de classement et de présentation des résultats : catégorisation, réseaux sémantiques, analyse de contenu**…

Bien que ces trois domaines soient fortement imbriqués et interdépendants, il est nécessaire de les distinguer, à la fois conceptuellement et pratiquement. Tout d'abord, ils renvoient à des phases différentes du fonctionnement des outils : schématiquement l'indexation, le traitement de la requête, le classement des résultats, le traitement et la présentation des résultats. Par ailleurs, ils traduisent souvent des orientations technologiques spécifiques des outils : autrement dit, les outils de recherche se distinguent de plus en plus selon la priorité qu'ils accordent, soit aux traitements linguistiques, soit aux fonctionnalités de recherche, soit aux techniques de traitement des résultats. Il est assez rare de trouver un moteur ou un métamoteur combinant toutes ces possibilités, qui paraissent s'exclure mutuellement (pour le moment du moins).

### Les différents niveaux d'analyse linguistique sur le web

Les documentalistes de CDI savent bien qu'avant de rechercher, il faut collecter et surtout indexer l'information. Les moteurs de recherche du web se présentent, sur ce plan, comme de gigantesques bases de données, dont les principes ne sont guère éloignés de ceux des logiciels de gestion documentaire. Mais avec des différences notables : la taille de leur index bien sûr, (6 milliards de documents pour Google), la nature des ressources du web, mais aussi (et surtout) les techniques de recherche sur le texte intégral, là où les logiciels documentaires « classiques » cherchent plutôt sur les métadonnées (catalogage/indexation) des documents.
De fait, les moteurs de recherche, utilisent des techniques d'indexation automatisée, issues de ce que l'ingénierie linguistique développe depuis une vingtaine d'années, et qu'on appelle les technologies du TAL ou du TALN (Traitement Automatisé des Langues ou Traitement Automatisé du Langage Naturel). Il faut noter que l'ingénierie linguistique (c.a.d. à la fois la recherche et les industries de la langue) s'est développée en dehors du web, proposant des solutions logicielles et des applications de plus en plus sophistiquées, pour le traitement et l'analyse linguistiques de corpus textuels limités, définis.
Sans approfondir, on peut relever au moins quatre niveaux d'analyse automatisée, correspondant aux quatre premières « couches » d'un texte : morphologique, lexicale, syntaxique, sémantique. Le cinquième niveau, celui de la pragmatique (le langage en actes, qui

[16] Par exemple Profusion**,** métamoteur spécialisé sur les ressources du web invisible. Disponible sur : <http://www.profusion.com>

permet à des humains de comprendre le sens implicite d'une phrase par son contexte), reste encore presque hors de portée de l'automatisation.[17]
A quels niveaux d'indexation se situent les moteurs de recherche ?
L'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.). Nous sommes là dans le domaine de l'analyse morphologique, fondée sur la seule reconnaissance de la forme des mots. Les flexions d'un verbe, les distinctions entre verbes et noms (« je porte la porte »), les synonymes, la polysémie, etc. : aucun piège du langage naturel n'est éliminé et ce premier niveau de l'indexation automatisée, aussi intéressant soit-il, génère un « bruit » considérable, expliquant le déluge des résultats sur les grands moteurs de recherche. Actuellement, la plupart des moteurs fonctionnent encore à ce **premier niveau de l'analyse morphologique**, parfois en éliminant les mots-vides (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)[18]. Concrètement, les mots au pluriel sont ramenés au singulier, les verbes sont mis à l'infinitif… Du coup, les index sont considérablement allégés, la recherche plus pertinente. Un moteur comme *Mirago* pratique ce type d'analyse linguistique[19]. L'analyse lexicale commence à se développer également sur plusieurs outils, avec les correcteurs orthographiques. 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 ».

Avec le troisième niveau d'analyse, nous passons **au stade de la grammaire, 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*[20], 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'usager, lui

---

[17] Une brève remarque pédagogique : au plan de la formation des élèves à la recherche d'information, s'il est important d'expliquer le fonctionnement des moteurs de recherche, on ne saurait réduire la formation aux seules techniques de requête et il y aurait ici des apprentissages multi-disciplinaires intéressants (entre français, linguistique et documentation) à développer, autour de ces niveaux d'indexation automatisée, des pièges du langage naturel, et des limites intrinsèques des moteurs de recherche…

[18] Bien que les spécialistes distinguent entre racinisation et lemmatisation, nous englobons ici sous le terme de lemmatisation tous les procédés « d'épuration » lexicale.

[19] Par exemple, si vous tapez « chevaux » sur Mirago, il ramènera également les pages contenant « cheval », contrairement à Google, qui ne ramènera que « chevaux ». Voir : <http://www.mirago.fr/>

[20] Voir http://www.exalead.com/cgi/exalead. Exalead équipe également la plate-forme de recherche d'AOL France : voir : http://www.aol.fr/

permettant d'affiner ses recherches. On trouve cette fonctionnalité sur quelques moteurs, comme *Alta Vista, Teoma, Voilà*, à des degrés différents et selon des technologies spécifiques.

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 encore peu répandu sur le web, et se rapproche le plus 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[21], semble-t-il, mais elle concerne surtout le traitement des résultats après une requête (cf plus loin) et non l'indexation *a priori* des documents. Un exemple intéressant de (véritable ?) 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. Le résultat, aussi problématique soit-il en termes journalistiques, est assez spectaculaire[22].

En résumé, l'intégration progressive des différents niveaux d'analyse linguistique et des méthodes de TAL dans le contexte du web reste, sans conteste, l'un des grands axes d'innovation pour les outils de recherche et certains moteurs, comme *Mirago* et surtout *Exalead*, mais aussi le moteur en langage naturel *AskJeeves*,[23] ont ouvert une voie prometteuse pour la recherche d'information, en dépassant la seule analyse morphologique.
Sans doute est-ce le chantier du Web sémantique qui généralisera ce quatrième niveau de l'indexation automatisée, celui qui intéresse le plus les humains que nous sommes, attachés d'abord à la signification des documents plus qu'à leur forme ou leur ressemblance…

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

Ce deuxième domaine d'innovations est plus simple à appréhender puisqu'il 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), 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. En effet, les moteurs[24] et certains métamoteurs[25] 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…),

---

[21] Par méthodes statistiques, on entend notamment le calcul des co-occurrences, c.a.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.

[22] Voir http://news.google.fr/

[23] Voir http://www.ask.com/

[24] 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. Voir : http://www.uhb.fr/urfist/Supports/ApprofMoteurs/ApprofMoteurs_cadre

[25] Comme Kartoo ou Ixquick : voir également le travail de comparaison de 6 métamoteurs que nous avons mené cette année : http://www.uhb.fr/urfist/Supports/ApprofMetamoteurs/ApprofMetamoteurs_cadre.htm

sur les pages similaires, sur différents champs des pages web (titre, liens, URL, métadonnées, etc.) et même sur la taille des fichiers (seulement sur AllTheWeb). La plupart de ces fonctionnalités de recherche, accessibles soit en mode simple, soit en mode avancé, restent généralement méconnues des utilisateurs, alors que leur connaissance et leur maîtrise sont l'une des conditions d'une recherche d'information efficace. La formation des élèves, qui s'en tiennent le plus souvent à un ou deux mots-clés en mode simple, ne consiste-t-elle pas, pour une part, à les initier à toutes ces possibilités et subtilités, permettant de trouver des documents très précis ?

**Catégorisation, réseaux sémantiques, analyse de contenu… : de nouveaux traitements de résultats**

Nous n'évoquerons pas, pour ne pas alourdir cet article déjà dense, les méthodes de classement des résultats, c.a.d. les principes selon lesquels les moteurs de recherche classent, ordonnent leurs résultats **: l'indice de pertinence** ou **l'indice de popularité**. Ces méthodes sont désormais assez bien établies[26] et l'innovation est désormais ailleurs : elle touche plutôt ce qu'il faut bien appeler, faute de mieux, **le traitement des résultats**, qui dépasse quelque peu les seules problématiques de classement.
Trois innovations importantes et intéressantes 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*[27], 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 (et la force) 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[28]. Concrètement, à partir de la requête « cheval de course », Exalead a généré, à partir des 68111 résultats, quatre grandes rubriques (Sport, Commerce et Economie, Régional, Sciences), avec des sous-rubriques (Elevage dans la rubrique Commerce et Economie).
Les documentalistes scolaires verront immédiatement l'intérêt pédagogique d'un outil comme Exalead, qui permet aux élèves, non seulement d'affiner leurs requêtes (notamment par la proposition de mots-clés, évoquée plus haut) mais aussi de percevoir de nouvelles pistes de recherche, de nouveaux liens vers d'autres thèmes, ainsi que l'environnement sémantique de leur sujet de recherche. En d'autres termes, les technologies de catégorisation des résultats

[26] Rappelons brièvement leurs principes :
- l'indice de pertinence permet de classer les documents selon les mots-clés du document (nombre, emplacement, « poids » des mots-clés….). Fondé sur de purs calculs statistiques, il a constitué la première méthode (et la plus utilisée) pour classer les résultats sur les moteurs de recherche. Il a été fortement bousculé par l'indice de popularité de Google.
- 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).

[27] Voir <http://vivisimo.com/>

[28] Pour mieux comprendre la différence entre classification « a priori » et « a posteriori », on peut prendre l'image des deux méthodes de classement d'une pile de livres : soit on les classe en fonction des rubriques d'une classification déjà établie (par exemple la Dewey), soit on part des livres eux-mêmes et on les regroupe selon leurs véritables thèmes et leurs ressemblances.

réintroduisent du sens, de la signification, de la structuration dans le chaos informationnel du 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*[29] et *MapStan*[30], 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 communs 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 des pages web ; les liens entre les nœuds représentent les relations entre les pages web, c.a.d. leur proximité, leur similarité.
Représentés sous forme de sphères et de liens sur *Kartoo*, de places et de rues sur *MapStan*, ces sortes de 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 précieuses sur l'importance de tel ou tel site (par le nombre de liens qu'il reçoit)… En bref, comme leur nom l'indique, ces métamoteurs graphiques développent, même de manière encore limitée, une nouvelle cartographie de l'information, permettant de mieux se représenter l'espace réticulaire du web.

Une troisième orientation technologique porte sur **l'analyse automatique du contenu des documents**. Elle est développée notamment par un métamoteur américain, *SurfWax*, particulièrement innovant[31]. 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 s'il existe, les mots-clés de la requête dans leur contexte, les points clés (*Key Points*) de la page, cad les phrases considérées par SurfWax comme les plus importantes. En bref, une sorte d'analyse des documents, permettant à l'utilisateur de mieux faire ses choix, d'affiner et d'élargir sa recherche.

On le voit : les innovations techniques ne manquent pas et offrent des perspectives tout à fait intéressantes pour les usagers, notamment pour les élèves confrontés au déluge et au chaos informationnel. Comme nous l'avons 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 catégorie de documents, de l'autre la représentation cartographique d'un réseau, ou encore l'analyse du contenu. Exclusives l'une de l'autre au plan technique (pour le moment du moins), disponibles surtout sur des métamoteurs (hormis Exalead), ces trois techniques représentent trois directions de recherche et d'innovation

[29] Voir <http://www.kartoo.com/>

[30] Voir <http://search.mapstan.net/>

[31] Voir 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.

appelées à se développer et à se perfectionner dans les années à venir. Pour les documentalistes de CDI et les élèves, ces technologies, et les outils qui les incarnent, devraient être mieux connues et surtout utilisées conjointement, selon les besoins et les contextes. Car s'il fallait tirer une leçon de cette brève présentation des outils et des technologies de recherche, ce serait sans nul doute cette sorte d'impératif catégorique de la recherche d'information : il faut utiliser (et inciter les élèves à utiliser) plusieurs outils, selon ses besoins et ses objectifs, et aussi pour sortir de la « googlemania » galopante, dangereuse comme toute pratique exclusive.[32]

## Et demain ? Vers le « Web sémantique » ?

On ne peut terminer un panorama de la recherche d'information sur Internet sans évoquer cette expression qui commence à faire florès, suscitant de plus en plus d'articles, de conférences, de formations… dans les métiers de l'information et qui, au-delà de l'inévitable effet de mode, peut représenter une mutation tout à fait majeure, non seulement de la recherche d'information, mais des usages du web : nous voulons parler ici du « Web sémantique ». Un tel sujet mériterait à lui seul un article complet et il ne peut s'agir ici que de fournir quelques éléments d'information, permettant une première approche d'une thématique particulièrement complexe. Nous essaierons de le faire à partir de questions simples.

**D'où vient et qui est derrière 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*). Rappelons que le W3C est un consortium créé en 1994, fondé sur 3 pôles de recherche internationaux (le MIT, la Keio University au Japon et un regroupement de 18 centres de recherche européens, ERCIM[33]), et regroupant de nombreuses entreprises informatiques, différents centres de recherche, etc., 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 (sous forme de recommandations), pour le maintien et l'évolution du World Wide Web[34].

**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 nombreuses limites et inconvénients du web, qui ont transformé celui-ci en véritable 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 hypertexte entre les sites web, bien que porteurs de sens pour les humains, n'ont aucune signification utilisable par les machines. Et pour bien comprendre le projet du web sémantique, il faut partir du constat, opéré

[32] Google serait-il à la recherche d'information ce que Microsoft est aux logiciels d'exploitation et d'application ?

[33] ERCIM : European Research Consortium for Informatics and Mathematics

[34] 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.

par le W3C lui-même, des diverses insuffisances du web ; citons par exemple : 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[35], les limites des outils de recherche, incapables encore d'analyser vraiment les pages web et de fournir la bonne information pertinente et adaptée aux besoins de l'utilisateur. Comme l'indique Philippe Laublet, l'un des chercheurs français impliqué dans le projet du Web sémantique, 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.* »[36]
On peut donc résumer ainsi l'objectif du projet de « *Semantic Web* » : pallier les insuffisances du Web en lui ajoutant une « couche » sémantique, permettant de libérer les utilisateurs d'une partie de leurs tâches de recherche afin de mieux exploiter l'information contenue dans les ressources du web. En fait, le web sémantique vise à mettre de l'ordre et surtout de la signification dans le chaos informationnel, en développant des outils et des méthodes qui rappelleront quelque chose aux documentalistes et aux bibliothécaires. Car il s'agit surtout de pouvoir identifier, décrire et indexer les innombrables 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 particulièrement complexe de plusieurs « couches » de langages et d'applications informatiques, plus ou moins autonomes. Très schématiquement, on peut relever au moins **quatre « couches »,** constituant autant d'axes de recherche complémentaires : l'identification, la structuration, la description et la représentation des ressources.

- **L'identification précise des ressources : les URI**

Pour pouvoir les décrire, les combiner, les associer et les utiliser, encore faut-il que chaque ressource électronique (texte, image, son…) soit identifiée de manière univoque ; 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*)[37]. Ce « méta-langage » (car 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) est fondé 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…). XML est actuellement en passe de se généraliser sur le web et de devenir le principal format d'échange des documents, et ce de manière transparente pour l'utilisateur. Avec ce passage de HTML à XML sur le Web, c'est une nouvelle

[35] 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.

[36] P. Laublet, *Introduction au web sémantique*, support Power Point, Journée de formation à l'URFIST de Rennes, 26 mai 2004

[37] Pour comprendre ce qu'est XML, voir : « Le XML expliqué à vos enfants ». *Archimag*, n° 159, novembre 2002

mutation silencieuse du document électronique qui s'opère sous nos yeux, que l'on pourrait résumer comme le passage de la forme au contenu. Les conséquences en seront immenses, même si elles restent encore difficilement cernables. Et bien que XML ait été développé indépendamment du Web sémantique[38] il en est néanmoins l'un des fondements techniques.

- **Une description complète, structurée et pertinente des ressources** : **les métadonnées, le RDF.**

Les métadonnées ne sont rien d'autre que des données à propos d'autres données. Si le principe est très ancien, puisqu'il est à la base du catalogage et de l'indexation, autrement dit de la description d'un document, le contexte du web et du document numérique change évidemment la donne et 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*[39], la *TEI*[40], *l'EAD*[41]…). Il existe une grande variété de systèmes et de standards de métadonnées et le Web sémantique, par rapport à cet univers foisonnant et hétérogène, 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 ici le **RDF** (*Resource Description Framework*) : il s'agit, non d'un nouveau format de métadonnées, mais d'un métalangage (sorte de méta-métadonnées !), offrant une syntaxe universelle qui permettra aux machines d'échanger des informations de métadonnées incompatibles. Sans entrer dans des explications techniques qui nous dépassent totalement, on peut tenter de cerner le principe de base du RDF, celui du « triplet » logique. RDF distingue trois types d'éléments : un sujet, une propriété, un objet, ou encore une ressource, une propriété, une valeur.

Par exemple, à propos du CNDP et de SavoirsCDI, on pourrait établir deux triplets, selon ce modèle[42] :

- d'une part, <CNDP><éditeur><SavoirsCDI> : le « sujet » CNDP, en tant qu'il a pour « propriété » d'être éditeur, publie un « objet », une « valeur » appelée SavoirsCDI
- d'autre part, <CNDP><statut><établissement Education Nationale> : le CNDP a, pour sa propriété « statut », la qualité d'être un établissement de l'Education Nationale.

Le croisement des deux triplets permettrait à une machine d'inférer que SavoirsCDI est une publication d'un organisme du MEN, même si SavoirsCDI n'est pas référencé comme tel. Et lors d'une recherche sur toutes les publications du Ministère de l'Education Nationale, on pourrait donc obtenir la référence de SavoirsCDI, sans le connaître au préalable.

Ce petit exemple illustre assez bien les potentialités du Web sémantique dans la recherche d'information. Même si ce projet relève encore en partie du rêve ou de la science-fiction, on

---

[38] Mais également dans le cadre du W3C. Pour un bref historique et une présentation simplifiée de XML, voir, dans notre support de formation : « *Recherche d'information sur Internet : approfondissement* », la partie sur le Web sémantique et sur XML : http://www.uhb.fr/urfist/Supports/Rechinfo2/Rechinfo2_cadre.htm

[39] Le Dublin Core : système de métadonnées élaboré en 1995 avec la participation de bibliothécaires, permettant de décrire une grande variété de ressources sur internet, à partir d'un ensemble de 15 rubriques de description.

[40] 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.

[41] L'EAD (*Encoded Archival Description*), conçue en 1993, vient du monde des archivistes et des bibliothécaires américains et permet de décrire très finement les fonds d'archives.

[42] Exemple emprunté à *La Dépêche du GFII*, n° 471, 4 mars 2004.

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.**

Si la définition exacte de ce terme, emprunté par l'informatique à la philosophie[43], reste souvent incertaine dans le contexte informatique, il importe d'en saisir l'objet et les caractéristiques. 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) : par exemple, un document scientifique *est un* type de document (il appartient à la catégorie « document »). 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, consensuelle (obtenue évidemment au terme de longs processus de discussions) de leur champ et des concepts qui le constituent, sur leur définition et sur les relations entre les notions. Une ontologie est véritablement une « vue sur le monde », ni vraie ni fausse, mais opératoire, partagée et, bien sûr, utilisable par les machines. On en comprend alors la fonction dans le Web sémantique : les ontologies y jouent le même rôle que les classifications, les thésaurus et autres langages documentaires dans les bibliothèques. Et ce rôle est essentiel, puisqu'il s'agit de permettre aux machines d'établir les liens sémantiques entre différentes ressources.

Si l'on reprend notre exemple de SavoirsCDI, quelle serait la condition nécessaire pour trouver SavoirsCDI dans une recherche sur toutes les publications du Ministère de l'Education Nationale ? Il faudrait une ontologie, permettant à la machine de valider les inférences mises en jeu dans les « triplets » évoqués plus haut, du type : « *Si* le CNDP *est un* établissement de l'Education Nationale et que le CNDP *est* l'éditeur de SavoirsCDI, *alors* SavoirsCDI appartient à la catégorie des publications du MEN. »[44].

Sur le web, sémantique ou non, co-existent plusieurs types d'ontologies, plusieurs outils et formats permettant de les créer, et il semble qu'une nouvelle étape a été franchie avec la publication récente, par le W3C, de la Recommandation sur le format **OWL** (*Ontology Web Language*), qui fournit un nouveau standard permettant d'homogénéiser l'élaboration de ces ontologies.

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

Rassurons-nous : si les fondements techniques du Web sémantique sont complexes et guère faciles à appréhender, leur usage sera transparent pour l'usager. Mais surtout, ils 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

[43] 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).

[44] Exemple emprunté à *La Dépêche du GFII*, n° 471, 4 mars 2004.

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. La recherche d'information sortira, enfin !, de la seule recherche morphologique, par comparaison de chaînes de caractères, pour aborder la « recherche intelligente » sur les contenus. Le Web sémantique permettra également de nouvelles possibilités de gestion des contenus par les gestionnaires de sites, des combinaisons infinies de données provenant de différentes sources, compatibles entre elles grâce à « l'interopérabilité » des standards RDF et OWL. Il devrait enrichir également les « Web Services », existant déjà en informatique d'entreprise : les « web services » offrent aux clients une grande variété de services fonctionnels, fusionnés sur un même système d'information comme, par exemple, dans le domaine du tourisme, où il est déjà possible, sur certains sites, de procéder de manière unifiée et transparente à la réservation d'avion, d'hôtel, de voiture, de visites guidées, etc.
En bref, le Web sémantique devrait permettre de surmonter l'hétérogénéité actuelle des ressources du web, souvent rédhibitoire pour l'internaute, et d'intégrer ces ressources sur une même interface, à partir d'outils simples à utiliser.

## Et les documentalistes dans tout cela ?

Nous l'avions dit en introduction de ce dossier : la recherche d'information sur Internet est difficile à cerner et les évolutions y sont particulièrement rapides. Nous n'en avons montré ici que les principales, en essayant d'en expliquer les principes techniques. Quant à la question des usages, potentiels ou réels, tant des outils de recherche que de l'information par les usagers, elle reste largement ouverte. Pour conclure ce dossier, il nous faut évoquer les conséquences de ces évolutions pour les documentalistes de CDI.
A nos yeux, deux questions majeures, déjà posées aujourd'hui, sont appelées à s'exacerber, à devenir de plus en plus vives, questions auxquelles nous ne répondrons pas ici mais qu'il semble important de rappeler et de débattre :

- **la question de la formation des élèves aux outils de recherche :**
  à quoi les former ? aujourd'hui, avec des outils de plus en plus simples à utiliser, demain, avec le web sémantique qui déchargera les usagers d'une bonne part du travail de recherche ? A partir du moment où les procédures techniques de la recherche d'information deviennent transparentes, où les interfaces de recherche sont intuitives et où l'usager est pris par la main par l'outil technique, la (vieille) question du contenu didactique de la formation des élèves se (re)pose toujours, selon l'alternative suivante : former aux procédures ou aux principes ? à l'utilisation pratique des outils ou à la compréhension de leur fonctionnement ? Le problème n'est pas nouveau et se pose aujourd'hui tous les jours dans les CDI et les bibliothèques : face à des élèves qui pratiquent tout seul *Google* (même fort mal), qui jonglent avec les outils de *peer-to-peer* pour télécharger des fichiers MP3, que peuvent apporter les documentalistes (hormis la sélection et le repérage des sources pertinentes) ? Est-ce qu'il ne faut pas développer la maîtrise intellectuelle, conceptuelle, des outils, par exemple expliquer aux élèves les principes élémentaires de fonctionnement d'un moteur de recherche, ce qui distingue un moteur d'un métamoteur, la différence entre une banque de données et un annuaire, etc. ? Face à un univers informationnel chaotique, illimité, confus et mouvant, il me semble que le rôle principal des professionnels de l'information est d'essayer d'apporter avant tout de la clarté, du sens et non de former des as de la procédure. Cela passe par un détour vers l'abstraction, à décliner et adapter bien évidemment selon les niveaux. Derrière cette question récurrente de la pédagogie de

l'information[45] se profile la question des invariants, des principes abstraits, universels de la recherche et de la gestion de l'information, autrement dit de cette « culture de l'information » qui se fait toujours attendre dans l'enseignement.

- **La question de la formation au questionnement et à l'évaluation :**

avec les progrès des outils de recherche et du web sémantique, qui allègeront les tâches procédurales de la recherche elle-même, s'opère un double déplacement des compétences informationnelles nécessaires : en amont et en aval de la recherche. Autrement dit, les étapes non automatisables de la recherche documentaire, à savoir d'une part le questionnement du sujet, l'élaboration de la problématique de recherche, d'autre part le filtrage, l'évaluation et la sélection de l'information, vont prendre un nouveau relief, au fur et à mesure des nouveaux progrès de l'automatisation. Tout le monde voit bien aujourd'hui que la question centrale, pour les élèves, n'est pas dans l'utilisation des outils de recherche, mais dans les capacités à identifier et évaluer des documents fiables et pertinents. Cette question, immense et complexe, de l'évaluation de l'information, dépasse de loin le cadre de cet article, mais on ne peut pas ne pas souligner le double lien entre les progrès des outils de recherche et le problème de l'évaluation : si certaines innovations techniques peuvent indéniablement aider l'usager à mieux repérer des documents pertinents (cf par exemple la catégorisation sur *Exalead*), si la spécialisation de certains outils peut contribuer à « assainir » les ressources proposées (en éliminant toutes les scories des moteurs généralistes), aucun outil, aucune innovation technique ne pourra remplacer l'usager, seul capable de décider si tel document est pertinent ou non par rapport à son questionnement. La question de l'évaluation de l'information, souvent simplifiée et réduite à des « recettes » techniques, repose en fait la part irréductible de « l'humain » dans les processus informationnels, la dimension socialisée de l'information, produit d'une interaction entre un document et un regard humain. Les documentalistes, souvent bousculés par les innovations techniques qui automatisent leur savoir-faire, gardent (et garderont longtemps !) ici toute leur importance, leur rôle-clé : celui d'être des médiateurs de l'information.

## INDICATIONS BIBLIO- (et Webo) GRAPHIQUES

**- ADBS. Journée d'étude *"Du thésaurus au web sémantique : les langages documentaires ont-ils encore un avenir ?"*, 11 avril 2002, Paris-La Défense. Document des pré-actes de la Journée.**

**- ANDRIEU, Olivier. Revue de presse "*Actu Moteurs*" [en ligne]. Site Abondance.com. Disponible sur WWW : <http://www.abondance.com>**

[45] Je me permets de citer un texte personnel : Alexandre Serres, *La triple dialectique des contenus de formation à la maîtrise de l'information.* Assises Nationales pour l'éducation à l'information. Paris, 11, 12 mars 2003. [en ligne]. Paris : URFIST, 2003. Disponible sur : http://www.ccr.jussieu.fr/urfist/Assises/Ass-index.htm

**- CATELLIN, Sylvie. Sérendipité, abduction et recherche sur Internet. In *Emergences et continuité dans les recherches en information et communication*, Actes du XIIe Congrès national des SIC, UNESCO (Paris), 10-13 janvier 2001. Paris : SFSIC, 2001**

**- ERTZSCHEID, Olivier, GALLEZOT, Gabriel. *Chercher faux et trouver juste, Sérendipité et recherche d'information.* Congrès de la SFSIC, Bucarest, Juillet 2003. Disponible sur : http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/06/89/sic_00000689_02/sic_00000689.html**

**- FOENIX-RIOU, Béatrice. *Recherche et veille sur le web visible et invisible. Agents intelligents, Annuaires sélectifs, Interfaces des grands serveurs, Portails thématiques.* Paris : Bases, Ed. TEC&DOC, 2001**

**- INRIA. *La recherche d'information sur les réseaux. Cours INRIA, 30 septembre - 4 octobre 2002, Le Bono (Morbihan).* Paris : ADBS, 2002**

**- LARDY, Jean-Pierre. *Recherche d'information sur Internet. Méthodes et outils.* 7e éd. Paris : ADBS, 2001.**

**- LAUBLET, Philippe. *Introduction au Web sémantique.* Rennes : URFIST, 2004. Support de formation (sous Power Point) pour un stage URFIST, 26 mai 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**

**- LINK-PEZET, Jo. Echo des stages : Exalead. *La Lettre de l'URFIST de Toulouse*, n° 33, Disponible sur WWW : <http://www.urfist.cict.fr/lettres/lettre33/lettre33-61.htm>**

**- REMIZE, Michel. Recherche et gestion de l'information : convergence vers le métier documentaire. *Archimag*, n° 172, mars 2004, p. 44.**

**- Revue "*Netsources*". Paris : Bases Publications**

**- SERRES, Alexandre. *Sélection de ressources sur les outils de recherche*. Rennes : URFIST de Bretagne-Pays de Loire, 2003. (dernière mise à jour : 23 mars 2004) Disponible sur :** http://www.uhb.fr/urfist/Supports/ApprofMoteurs_Ressources.htm