Moteurs de recherche : des enjeux d'aujourd'hui aux moteurs de demain.

Olivier Ertzscheid

MOTEURS DE RECHERCHE :
DES ENJEUX D'AUJOURD'HUI AUX MOTEURS DE DEMAIN.

RESUME : Les moteurs de recherche occupent une place prépondérante dans nos accès à l'information et à la connaissance. Ils suscitent également de vives interrogations, notamment dans leur capacité à rendre indexable des informations relevant indistinctement des sphères publiques, privées et intimes des contenus disponibles en ligne. A l'heure où le modèle économique de ces outils semble stabilisé, nous nous efforcerons, au travers d'une mise en perspective de leurs principaux développements technologiques, d'une analyse des pratiques en recherche d'information, et d'un état de l'art des questionnements sociétaux actuels, de fournir quelques bases pour une analyse prospective de ce secteur. Une première partie s'attachera à décrire la constitution des moteurs de recherche tels que nous les connaissons aujourd'hui, comme autant de machines sociales. Dans un second point nous reviendrons sur les développements et directions actuellement les plus significatifs, permettant d'atteindre le graal d'une « recherche universelle ». Dans une troisième partie nous présenterons les enjeux qui, dès demain, seront déterminants pour les moteurs de recherche et la perception qu'ils nous renvoient du web.

INTRODUCTION

J'ai 10 ans.

En septembre 2007 le moteur Google fêtait ses 10 ans d'existence. En Juin 2008, l'annuaire Dmoz (projet OpenDirectory) soufflait également dix bougies. A cette dernière exception près, le modèle des annuaires de recherche a aujourd'hui atteint ses limites, dans l'incapacité de s'aligner sur l'explosion de la masse documentaire à l'échelle du web. Pour les moteurs en revanche, plus précisément pour les quelques grands acteurs qui dominent ce secteur, l'avenir apparaît radieux. La multitude des services qu'ils développent et proposent – généralement gratuitement – à l'usager, l'impérieuse nécessité de leur usage pour qui veut accéder aux informations et connaissances aujourd'hui disponibles sur le réseau, les rend tout simplement incontournables. En 2008, leur modèle économique est constitué (publicité « contextuelle et ciblée » et liens « sponsorisés ») et génère d'énormes profits. Pour autant, cet âge d'or n'est pas nécessairement pérenne et présente quelques risques pour des usagers trop candides.

La plasticité du web, la dynamique de ses acteurs, les nouveaux entrants sur ce secteur et le nombre des chantiers technologiques actuellement en cours et restant à mettre en place rend cependant très délicat tout exercice de prospective. Cet article s'efforcera donc d'observer la situation actuelle des acteurs de la recherche d'information sur Internet, de pointer et d'analyser les changements technologiques en cours à la lumière des usages émergents ou déjà actés, et de soulever un certain nombre de questionnements qui font aujourd'hui problème. Des questionnements principalement liés à la position hégémonique de certains moteurs, à leur capacité à capter et à monétiser des pans entiers de nos activités et de nos vies numériques. Soit le passage du World Wide Web au « Giant Global Graph » pour reprendre l'expression de Tim Berners Lee (BERN, 2007).

Giant Global Graph ?

L'expression fait référence à trois niveaux d'analyse différents mais déterminants pour comprendre la genèse et l'avenir des moteurs de recherche en particulier et des acteurs de

l'accès à l'information et de la diffusion de contenus en général. Ces trois niveaux sont ceux du « Net » en tant qu'infrastructure technologique, du « Web » comme plateforme de mise à disposition des contenus, et du « Graph » désignant la capacité à produire dynamiquement de l'information, des connaissances, sur la base d'un recoupement des informations et connaissances déjà existantes. La notion de « graphe » était déjà présente dans la vision initiale de Tim Berners Lee quand celui-ci « inventa » le web actuel (BERN, 1989). Mais elle prend aujourd'hui une tout autre dimension grâce aux travaux autour du web sémantique (BERN 2001) et à l'émergence et l'engouement massif suscité par les réseaux sociaux. Sans anticiper sur les conclusions de cet article, il apparaît aujourd'hui très clairement que c'est sur leur capacité à articuler ces deux dimensions (sociale et sémantique) que seront construits les moteurs de demain.

1. DES MACHINES SOCIALES

La réalité sociale du web compte aujourd'hui près d'un milliard d'internautes (NIEL, 2005). Sa réalité documentaire fait état d'estimations autour de 30 milliards de pages web[1] en Février 2007, alors que la dernière version – 2003 – de l'étude quantitative systématique menée à Berkeley (LYMA, 2003) faisait état d'un web « de surface » – c'est à dire indexé ou indexable par les moteurs – représentant 167 térabits d'information. Au vu de cette seule volumétrie et de ces chiffres qui confinent à l'incommensurable, on comprend mieux l'impérieuse nécessité de l'existence d'outils de repérage, de classification et d'accès. De fait, les moteurs de recherche ont vu le jour dès que le nombre de serveurs web est devenu ingérable à l'aide de simples signets ou de listes d'adresses. La démesure de ce continent documentaire a ensuite rapidement conditionnée les logiques de déploiement des outils permettant de s'y retrouver : le repérage et l'accès prenant le pas sur la classification ordonnée.

1.1. Description, restitution, prescription.

Description

Les premiers temps du web virent d'abord la reconduite d'un schéma de traitement documentaire « classique », c'est à dire la sélection, la validation et le classement thématique des sites web par des individus. Yahoo !, premier annuaire historique, était au départ une simple liste de signets, maintenue par ses deux fondateurs, chaque site retenu étant accompagné d'une **description** de son contenu. Décrire et hiérarchiser thématiquement des contenus validés étant les deux principales caractéristiques de ces annuaires.

Restitution

Compte tenu de l'explosion volumétrique déjà mentionnée, et grâce aux progrès de l'ingénierie linguistique permettant l'indexation rapide de très larges corpus, les moteurs de recherche firent ensuite progressivement leur apparition, dotés de « spiders »[2] leur permettant d'atteindre rapidement les sommets d'une échelle quantitative autrement ingérable. Dans l'imaginaire de l'utilisateur comme dans la réalité technologique, l'avantage est au moteur disposant de la plus grosse base d'index. Cette course au quantitatif occasionna pendant

[1] http://www.boutell.com/newfaq/misc/sizeofweb.html

[2] Le spider ou robot d'indexation est un logiciel parcourant le web de liens en liens afin d'indexer les sites visités et de mettre à jour les bases de données des moteurs de recherche.

quelques temps une « guerre d'affichage » sur les pages d'accueil des moteurs, guerre dans laquelle entrait une bonne part de marketing. Cette stratégie d'affichage est aujourd'hui abandonnée par la plupart des acteurs[3] et il est donc très difficile d'obtenir une vue non-biaisée du volume des contenus effectivement indexés. La seule certitude est que ceux-ci se comptent en milliards (de pages, d'images …). L'urgence à laquelle permirent de faire face les moteurs de recherche fut celle de la **restitution** en temps quasi-réel[4] du rythme de publication propre au web.

Prescription.

La bataille actuellement engagée et celle de la **prescription**. A l'heure ou la plupart des contenus, quelle que soit leur nature, sont presque instantanément disponibles, visibles dans les moteurs, et devant le gigantisme des mêmes contenus et l'étroitesse de la vue qui nous en est proposée par les moteurs et par nos usages[5], l'internaute a un besoin croissant de prescription et de conseil. Il s'agit de disposer d'informations supplémentaires permettant de réduire l'écart entre les contenus effectivement disponibles et pertinents et ceux renvoyés par les moteurs[6] ainsi que la part d'aléatoire inhérente au choix des mots-clés déposés par l'internaute. Cette prescription peut prendre plusieurs formes. Les sites marchands (Amazon par exemple) proposent ainsi divers systèmes plus ou moins fiables[7] de « recommandation »[8] permettant, sur la base d'un acte d'achat initial, d'orienter les internautes vers des produits semblables ou similaires. De leur côté les moteurs de recherche mettent en pratique différentes stratégies. Il est aujourd'hui possible de créer des moteurs personnalisés[9] sur la base d'un corpus de sites web défini par l'internaute. Une autre solution consiste à valoriser la capacité de prescription affinitaire des « proches », des « amis » ou des « collaborateurs » de l'internaute. Moteurs de recherche et réseaux sociaux travaillent ainsi à des rapprochements qui permettraient d'optimiser la capacité de restitution des premiers et la force de prescription des seconds.

1.2. Algorithmes sous le moteur.

Les algorithmes d'indexation et de classement constituent le cœur technologique des moteurs de recherche. Celui de Google, baptisé « Pagerank » constitua à l'époque de sa mise en œuvre une véritable (r)évolution technologique, aujourd'hui copiée par l'ensemble des acteurs majeurs du secteur. L'heure est à la convergence des algorithmes de pertinence même si, comme nous le verrons plus loin, les tentatives de diversification et de différentiation sur ce secteur technologique restent importantes. Rappelons d'abord que le PageRank est inspiré des indicateurs scientométriques définis par Eugène Garfield (GARF, 1972) pour l'évaluation des publications scientifiques. En lieu et place du nombre de citations d'un article scientifique dans la bibliographie d'autres articles, la pertinence d'une page web est définie à partir du

---

[3] Sur ce sujet voir notamment (ERTZ, 2005a) et (VERO, 2005).

[4] Le temps de prise en compte des sites dans l'index des moteurs, s'il s'améliore constamment, reste cependant variable. En revanche, pour des contenus de type actualité (news) ou carnets web (weblogs), cette indexation se fait quasiment en temps réel, amenant à parler d'un World **LIVE** Web.

[5] La plupart des internautes ne vont pas au-delà de la consultation de la première page de résultats.

[6] Il s'agit du « taux de rappel », c'est à dire le ratio entre le nombre de documents pertinents trouvés lors d'une recherche et le nombre total de documents pertinents existant dans le système. L'autre indicateur est le « taux de précision » (ratio entre le nombre de documents pertinents trouvés lors d'une recherche et le nombre total de documents trouvés en réponse à la question).

[7] Voir (DAVI, 2006)

[8] pour une vue plus complète de ces systèmes de recommandation, voir (ERTZ, 2008b)

[9] Rollyo (http://rollyo.com), Eurekster (http://www.eurekster.com/), Yahoo ! Search Builder (http://builder.search.yahoo.com/m/promo), Google Custom Search Engine (http://www.google.com/coop/cse/)

nombre et de la pertinence des pages qui la citent (backlinks). C'est une mesure quantitative qui est utilisée pour construire des métriques qualitatives. Le « coup de génie » des fondateurs de Google fut de parvenir à transposer ce mode opératoire depuis un corpus très spécifique et normé et disposant de puissants filtres éditoriaux en amont (publications et édition scientifique) vers un corpus ouvert, non filtré et généraliste : le web.

Contrairement à une première idée reçue, le brevet du Pagerank originel déposé à l'époque ou Serguei Brin et Larry Page étaient étudiants à l'université de Stanford, est public[10] et l'ensemble des documents originaux publiés par ses inventeurs est disponible (BRIN, 1998a, 1998b, 1999). On dispose également d'un grand nombre d'analyses extérieures décrivant et analysant son fonctionnement (LANG, 2003). Le fonctionnement du PageRank originel est donc connu, ce qui permit aux autres acteurs de s'en inspirer et de faire de l'analyse des backlinks l'un des tous premiers critères pour l'affichage de leurs listes de résultats et la constitution de leur base d'index. En revanche, la domination technologique de Google est liée à la difficulté d'évaluer le poids exact de l'indicateur relationnel dans l'algorithme de pertinence. Dans un billet en date du 20 Mai 2008 sur le blog officiel de Google[11], on apprend que pour l'année 2007 le Pagerank a connu plus de 450 modifications ! Un enrichissement naturellement confidentiel mais qui passe notamment par la prise en compte de logiques qualitatives en amont (nature et sémantique des backlinks), par de puissantes procédures de filtrages en aval (permettant d'atténuer les usages détournés[12]), et par une infrastructure technique permettant de travailler à une échelle statistique et computationnelle hors-norme[13].

Fièvre algorithmique

Les algorithmes de l'ensemble des moteurs sont de la même manière en permanence modifiés. De plus, certains types de contenus peuvent nécessiter une algorithmie dédiée. C'est notamment le cas de l'indexation image pour laquelle Google vient de publier une approche différente de celles couramment utilisées : le VisualRank (JING, 2008). L'indexation image est un secteur stratégique essentiel pour le futur de la recherche d'information, particulièrement dans la tendance actuelle d'une recherche globalisée et universelle[14].

Complexité algorithmique objectivée ou panoptique subjectif ?

Quand nous consultons une page de résultat de Google ou de tout autre moteur, nous ne disposons pas simplement du résultat d'un croisement combinatoire binaire entre des pages répondant à la requête et d'autres n'y répondant pas ou moins (*matching*). Nous disposons d'une vue sur le monde (*watching*) dont la neutralité est clairement absente. Derrière la liste de ces résultats se donnent à lire des principes de classification et d'organisation de l'information et des connaissances : l'affichage lisible d'une liste de résultats, est le résultat de l'itération de principes non plus seulement implicites (comme les plans de classement ou les langages documentaires utilisés dans les bibliothèques) mais invisibles et surtout dynamiques, le classement de la liste répondant à la requête étant susceptible d'évoluer en interaction avec le nombre et le type de requêtes ainsi qu'en interaction avec le renforcement (ou l'effacement) des liens pointant vers les pages présentées dans la page de résultat (ERTZ, 2004).

L'autre grande « nouveauté » qu'apportent les moteurs de recherche au mode de

---

[10] http://tinyurl.com/yswq4g

[11] http://googleblog.blogspot.com/2008/05/introduction-to-google-search-quality.html

[12] le « spamdexing » notamment, un référencement frauduleux visant à tromper l'indexation par les moteurs grâce à de faux mots-clés.

[13] On parle actuellement de « cloud computing » pour désigner le recours à des infrastructures informatiques largement distribuées.

[14] cf la seconde partie de cet article.

circulation, d'organisation et d'accès habituel aux connaissances est l'assujetissement à des logiques marchandes dans lesquelles la publicité EST du contenu[15].

1.3. Concentré d'économies et économie concentrée ...

La caractéristique principale de la toile mondiale n'est pas tant qu'elle a permis de rendre disponibles des milliards d'informations, mais surtout qu'elle a amené des millions d'utilisateurs à faire de la recherche d'information et de la recherche de documents une tâche quotidienne. Dans cette tâche, les moteurs de recherche en général et Google en particulier, sont généralement les seuls intermédiaires entre l'expression d'un besoin et sa partielle ou totale satisfaction.

En France, le moteur Google représente à lui seul près de 75% de l'ensemble des recherches effectuées[16]. Cette proportion pour les Etats-Unis, bien que plus faible (50%)[17], marque cependant une claire domination du moteur, qui se mesure autant à l'aune des parts de marché qu'il représente, qu'à son adoption généralisée, avec par exemple le passage dans l'usage de l'expression « Googler quelque chose ou quelqu'un » qui désigne métonymiquement l'ensemble du processus de recherche sur Internet.

Le service offert par les moteurs de recherche est un service gratuit. Mais le coût qu'il représente est maintenant considérable, aussi bien en termes de recherche et développement qu'en termes d'infrastructures. Les moteurs disposent aujourd'hui d'un modèle économique aussi rentable qu'efficace, l'affichage de bandeaux et de bannières publicitaires sur leurs pages d'accueil s'étant rapidement révélé insuffisant. L'arrivée de la publicité sous forme de liens sponsorisés permit aux moteurs de dégager des bénéfices à la hauteur des parts de marché qu'ils occupent[18]. Leur modèle – tout au moins pour des sociétés dans lesquelles la « recherche » est le cœur de métier[19] - est donc celui d'une régie publicitaire qui diffuse dans un panel de médias de plus en plus large (web bien sûr mais également plus récemment télévision et radio). La structure de l'offre est assez simple : il est possible d'acheter des mots-clés pour un affichage en première page de résultats, ou bien d'afficher sur son site, les liens sponsorisés contextuels envoyés par le moteur qui vous reverse alors un pourcentage pour chaque clic (offres Google Adwords et Adsense).

Le marché du « Search » comme l'appelle les américains est ainsi outrageusement dominé par Google, avec, très loin derrière lui Yahoo ! et Microsoft, en France comme outre-atlantique. La très récente proposition de rachat adressée à Yahoo ! par Microsoft (CROS 2008) a finalement été rejetée, l'occasion pour Google d'asseoir, *in fine*, encore un peu plus sa position en contractant avec Yahoo! un partenariat qui lui permettra de placer des publicités à côté des recherches menées par les internautes sur le moteur de recherche de Yahoo!

1.4. Une diversité de services ... au service d'une confusion des pratiques.

Si la publicité est la clé du modèle économique des moteurs, le corrélat est de réussir à augmenter au maximum le trafic sur leurs sites ou ceux de leurs services. D'où une stratégie de diversification desdits services et le virage vers une offre de type « portail » pour optimiser

---

[15] « Ads are content » selon la formule d'Omid Kordestani, en charge des programmes d'affiliation publicitaire chez Google (http://battellemedia.com/archives/001974.php)

[16] Source : IT Facts http://blogs.zdnet.com/ITFacts/

[17] http://searchenginewatch.com/showPage.html?page=3625336

[18] pour le 4ème trimestre 2007 les résultats financiers de Google étaient de 4,83 milliards de $ de revenus, soit une augmentation de 51 % par rapport à 2006 et 14% par rapport au 3ème trimestre. (Source : http://www.google.com/intl/en/press/pressrel/revenues_q108.html)

[19] ce qui est le cas pour Google et Yahoo! dans une moindre mesure, mais pas pour Microsoft qui reste d'abord un fabriquant de logiciels.

davantage la rentabilité publicitaire de ce public « captif ». Toute la logique de déploiement des services décrits ci-dessous correspond à un seul credo : maintenir le plus longtemps possible l'internaute dans la périphérie des services associés ou possédés par le moteur. Au-delà donc de la « recherche pure », les moteurs proposent aujourd'hui :

- de « communiquer » par l'envoi de courriels, le clavardage et des fonctionnalités de messagerie instantanée
- de « s'orienter » et de « géo-localiser » d'autres services grâce notamment à Google Earth[20]:
- de « chercher », non pas simplement sur leur page web, mais également par le biais de la téléphonie mobile, cette dernière constituant un secteur stratégique pour l'expansion et la monétisation de l'ensemble des services qu'ils proposent
- « d'organiser », de « s'organiser » et « d'indexer » : la plupart des services proposent des fonctionnalités de classement par dossiers et sous-dossiers, d'indexation à l'aide de mots-clés ou de tags. Pour l'ensemble des documents, des ressources, des produits, nombre de fonctionnalités de classement sont proposées.
- De « partager » tout type de document avec une liste de contacts choisis ou avec l'ensemble des utilisateurs desdits services : le passage à une offre de bureautique en ligne[21] fut la partie la plus visible d'une dérive annoncée des continents documentaires[22]
- De « créer » des blogs, des sites web, des albums photo, etc …
- « d'héberger » des contenus vidéos, iconograhiques …
- « d'analyser » les statistiques et logs de consultation de votre site[23]

Créer, héberger, rechercher, indexer, classer, s'orienter, localiser, partager, analyser … Cette liste d'actions repose sur une liste tout aussi conséquente (et non-exhaustive) de services soit créés soit rachetés par les moteurs : FlickR et Del.icio.us appartiennent désormais à Yahoo!, YouTube et Blogger sont la propriété de Google, pour ne citer que quelques-uns des sites emblématiques du web 2.0.

Dans cet écosystème, les usagers peuvent être rapidement désorientés ou ne pas prendre conscience des « filiations » entre ces différents services. Une situation qui satisfait pleinement les moteurs : plus leur écosystème sera large, moins les internautes auront la possibilité de s'en éloigner ou d'en sortir, plus la navigation se fera en son sein, et plus le modèle économique choisi – c'est à dire la publicité « en échange » de la gratuité – sera rentable.

1.5. Un moteur. Des recherches.

Cette confusion des pratiques s'ajoute à une diversité déjà ancienne des modes de requêtage en vigueur sur le web. (BROD, 2002) établit une taxonomie des recherches en fonction du « *besoin derrière la question* », distinguant ainsi trois classes :

- des recherches navigationnelles visant à retrouver une page particulière,
- des recherches informationnelles qui nécessitent la consultation de plusieurs pages
- des recherches transactionnelles enfin, qui manifestent un désir d'accomplir une action comme un achat en ligne.

---

[20] http://earth.google.com
[21] Office Live pour Microsoft (http://www.officelive.com/) et Google Docs (http://docs.google.com/)
[22] voir le point 2.1.
[23] Service Google Analytics : http://www.google.com/analytics

Au moment de l'étude (2002) la répartition entre ces requêtes est respectivement de 50%, 20% et 30%. Le panorama actuel des outils, conjugué à l'explosion des sites spécifiquement marchands et des comparateurs de prix, ainsi qu'à la levée progressive – en terme d'habitudes de consommation – des réticences à l'achat sur Internet, permettent de penser que la dernière catégorie doit aujourd'hui avoir connue une croissance significative.

Cette première sériation doit être affinée par la prise en compte de l'aspect de plus en plus communautaire de certains sites de recherche, qui se servent de la somme des expertises individuelles et/ou de l'analyse des actes d'achat individuels pour alimenter ensuite des systèmes de recommandation dont l'amplitude peut osciller entre le cercle plus ou moins large de pairs et/ou des amis et celui beaucoup plus large de l'ensemble de la communauté des utilisateurs de l'outil.

Ces recherches et ses actions s'inscrivent aujourd'hui dans un écosystème très large qui mêle indistinctement les univers informationnels publics, privés et intimes.

2. DERIVE DES CONTINENTS DOCUMENTAIRES ET RECHERCHE UNIVERSELLE.

Le web actuel se caractérise par une mixité et une perméabilité de plus en plus grande de ses contenus, phénomène que nous analyserons comme une dérive des continents documentaires. Les résultats de cette dérive sont tout à la fois sensibles dans les usages connectés qui viennent s'y greffer et dans la manière dont les acteurs de la recherche d'information proposent aujourd'hui des fonctionnalités de recherche « universelle » désignant la capacité pour l'utilisateur de chercher simultanément dans les différents index (et les différentes bases de données) proposés par les moteurs de recherche.

2.1. Dérive des continents documentaires

Première période

La représentation que nous avons du web est conditionnée par les possibilités de navigation et d'accès que proposent les moteurs de recherche. De sa naissance (BERN, 1989) jusqu'à la fin des années 1990 le web comme continent documentaire se confond avec le web public (« World Wide Web »), indexé par les moteurs. A ses côtés un web « profond » se constitue (BERG, 2001) : les pages sont générées dynamiquement à partir des requêtes déposées par les utilisateurs[24]. Les moteurs de recherche peinent encore (pour des raisons techniques) à indexer ces contenus, justifiant l'expression d'un « web invisible »[25]. En parallèle, l'échange de courriers électroniques et les documents stockés sur les ordinateurs personnels échappent à l'indexation des moteurs. La ligne frontière des continents documentaires visibles et invisibles tient donc à l'impossibilité d'accéder à certains types de contenus pour les indexer. Une frontière aujourd'hui abolie.

Deuxième période

Web public, web profond, correspondances électroniques personnelles mais aussi fichiers et documents stockés sur nos ordinateurs personnels sont désormais réunis en une même sphère d'indexabilité (ERTZ, 2005b). La raison : le passage « en ligne » de l'essentiel de nos comportements informationnels, grâce au déploiement d'outils dédiés mis à disposition

[24] par exemple sur des sites d'achat, de réservation ou sur des catalogues en ligne, etc.

[25] Le 11 Avril 2008, Google a annoncé qu'il allait être capable d'indexer certaines données situées "derrière" les formulaires web. Voir ce billet : http://affordance.typepad.com/mon_weblog/2008/04/de-profundis.html

par les moteurs de recherche (webmail, desktop search)[26]. L'essentiel du matériau documentaire qui définit notre rapport à l'information et à la connaissance se retrouve entre les mains de quelques sociétés marchandes : courriers privés, fichiers personnels, pages web publiques, pages web d'entreprises, publication savantes, fonds numérisés de bibliothèques. Un seul et même outil, une seule et même société commerciale[27] indexe et supervise l'accès à cet ensemble. En termes d'accès et de droit à l'information, l'extrême mouvement de concentration qui touche ici la médiasphère fait débat. D'autant que de nouveaux usages produisent une hybridation inédite des sphères publiques et privées : on parle à propos des blogs, d'espaces de publication « extimes » (TISS, 2001). Enfin, des comportements informationnels de plus en plus nomades se cristallisent autour des outils bureautique en-ligne offerts par les mêmes acteurs.

Soit un nouvel écosystème informationnel global préempté par quelques moteurs de recherche qui font commerce de l'accès à ces contenus. Rappelons ici que l'indexation massive de la sphère documentaire publique, privée et intime n'a plus comme objectif principal de répondre à des logiques de recherche d'information en optimisant la pertinence des résultats proposés. Elle vise la diffusion ciblée de publicités contextuelles sur tout type de contenu documentaire, dans tous les types d'activités sociales ou professionnelles connectées.

2.2. Le web comme base de données.

Sur la base de ce gigantesque continent réunifié, les moteurs vont progressivement mettre leur pratique d'indexeurs ainsi leurs algorithmes en cohérence avec cette nouvelle configuration des gisements, des « silos » informationnels. La première étape consistant à tenter de « rationaliser » la dimension profondément hétérarchique du web en s'inspirant de l'architecture et de la structuration des bases de données.

Pages statiques et web dynamique.

A ses débuts, le web était composé de pages dites « statiques », générées en HTML. Puis le développement de langages de programmation dédiés permit au web de devenir « dynamique ». La notion de page ou de web dynamique désigne les sites reposant sur une base de donnée permettant la composition et l'affichage de pages « à la volée », sur la base d'une requête déposée par l'internaute, par exemple au travers d'un formulaire de saisie ou de recherche. La plupart des sites de réservation (train, avion, hôtel, etc) et l'immense majorité des sites marchands fonctionnent sur ce système. Ce web dynamique compose également pour une grande part ce que l'on nomme le web invisible (cf supra) : les pages n'étant pas affichées en permanence mais générées à la demande et n'existant que le temps de la session de l'internaute, les moteurs de recherche n'ont pas la possibilité technique de les indexer. Sauf à « déporter » ces gisements d'information vers leurs propres serveurs, par exemple en proposant un service de création et de gestion … de base de donnée.

Dans la base des moteurs.

---

[26] Les webmails permettent de stocker et de consulter son courrier électronique en ligne. L'offre Desktop Search permet d'indexer le contenu d'un ordinateur personnel grâce au moteur choisi (Google, Yahoo ! ou Microsoft).

[27] Google dispose ici d'un leadership incontestable, lequel ne peut être élargi au delà des deux sociétés concurrentes que sont Yahoo et Microsoft.

Le 16 Novembre 2005, Google lance Google Base[28]. Le principe du service est très explicite : « *Tout le monde, des grandes entreprises aux gestionnaires de sites jusqu'aux individus peut soumettre son contenu sous forme de données ('data items'). Nous les hébergeons et les rendons accessibles gratuitement. (...) Cette version béta est un petit pas supplémentaire vers notre but, créer une bases de donnée en ligne d'information facilement accessible et structurée.* »[29] Comme dans une base de donnée « classique », chaque item peut être enrichi de divers attributs et de valeurs permettant d'y faciliter la recherche par champs. Il est également possible d'y ajouter des « labels » (dix au maximum) sous forme de phrases ou de mots-clés en complément des attributs, ainsi qu'une description (texte libre) et un contact (nom, numéro de téléphone, courriel …).

La logique de déploiement de ce service s'inscrit dans la continuité de l'ambition de Google et des autres moteurs : celle de rendre indexables et « cherchables » (et donc marchandisables) l'ensemble des produits, des biens, des informations et des connaissances. En se dotant de l'ossature architecturale d'une base de donnée à vocation planétaire, ces mêmes moteurs tracent les nouvelles frontières d'un continent documentaire réunifié. Ils se réservent du même coup la possibilité d'y apposer un jour des droits de douane dont ils seront les seuls bénéficiaires.

Ce projet et le format de structuration des données qu'il utilise (XML) fut perçu par certains analystes comme un pas significatif vers le web sémantique : « *dans le système Google base, les descripteurs sont à l'origine des entités du système, et collent au plus près aux besoins des utilisateurs. Il n'y a pas d'ontologie, mais un ensemble de labels permettant de créer des hiérarchies à la volée, complétés par des attributs. Si l'utilisation va dans ce sens, on pourrait considérer que Google base représente un outil de création de schémas d'annotation, de mise en place d'annotations/méta-données/items et de recherche dans ces items, soit une autre manière de se diriger vers un web « sémantique ».* »[30]

Bases de moteurs et flux de données

Le projet Google Base est révélateur d'une logique de délocalisation, de désintermédiation, de déplacement documentaire. Un déplacement qui affecte aujourd'hui l'ensemble du web, et qui est notamment perceptible derrière les sites emblématiques du web 2.0. « *L'explosion dont il est question concerne la bascule des contenus d'un site web d'une internalité à une externalité. Au lieu qu'un site web ne soit un « lieu » dans lequel les données « sont » et vers lequel d'autres sites « renvoient », un site web sera une source de données qui seront elles-mêmes dans de nombreuses bases de données externes, dont celle de Google (GoogleBase). Pourquoi alors « aller » sur un site web quand tout son contenu a déjà été absorbé et remixé dans un flux de données collectif ('collective datastream').* » (GREE, 2005)

Cette nouvelle externalité se donne particulièrement à voir dans les pages d'accueil personnalisables du type de celle de Netvibes[31]. Celles-ci ne comportent plus de contenu « interne » mais reposent sur une architecture informationnelle entièrement générée (et temporairement stabilisée, fixée numériquement) à partir de contenus informationnels tous externalisés[32]. Le contenu s'efface derrière l'architecture. Le discours n'est plus ancré dans un dispositif (technologique) mais le dispositif ancre le discours. Au delà de l'effervescence

---

[28] http://base.google.com

[29] http://googleblog.blogspot.com/2005/11/first-base.html

[30] Message de Yannick Prié sur la liste du RTP-DOC (http://rtp-doc.enssib.fr/)

[31] http://www.netvibes.com

[32] Il est possible d'y « amener » différents services et/ou informations comme la météo de ma région (piochée par exemple sur Yahoo), mon courrier électronique (capté par exemple dans Gmail), les fils RSS de presse extraits de mon agrégateur, et ainsi de suite.

technologique, il y a bien ici un changement de nature dans la forme du web, dans les modes d'agrégation, et dans l'instanciation de ses contenus.

Read Write Web

Sous l'impulsion des moteurs et avec l'assentiment des internautes qui voient dans ces usages un gain qualitatif incontestable, s'efface l'ancienne frontière entre la nature structurée – et donc à forte valeur ajoutée – des bases de données et celle profondément chaotique du Web. Un effacement à moindre frais, une partie significative des coûts de développement étant reportée sur l'intérêt des internautes à alimenter un tel service, intérêt s'expliquant lui-même par leur intérêt à voir leurs données dotées d'une visibilité inatteignable sans Google.

« *it's about the emergence of a data web made of loosely coupled sets of XML fragments that people and processes can easily read and write.* »[33] Le web se structure. Ce qui en soi n'est ni un bien ni un mal. Mais ce qui pose la question de la maîtrise d'œuvre de cette structuration et de ses finalités.

Le web devient essentiellement dynamique. Conditionné par des logiques de flux qui reposent de manière lancinante la question de l'archivage, celle de la trace.

Du point de vue de l'information perçue comme un écosystème, cette double dynamique caractérise le web 2.0, un web non plus simplement accessible en lecture mais également ouvert en écriture. Read/Write Web. C'est en tenant compte de ce paramètre que les moteurs déploient de nouvelles ambitions, de nouveaux services, de nouvelles stratégies.

2.3. La recherche universelle ou l'algorithmie ambiante ?

De la recherche universelle en particulier …

Le principe de la recherche universelle est donc de renvoyer sur la page de résultat des contenus en provenance de l'ensemble des bases proposées par le moteur (web mais aussi images, blogs, vidéos …). Ce renvoi peut prendre deux formes : les contenus différents sont directement insérés dans la page de résultat (« embed »), ou ils sont présentés sous forme de liens (« linked ») au bas de la page de résultats sur la base d'une requête identique ou augmentée d'un ou deux mots, lesquels liens pointent en fait vers d'autres services (images, news, maps …).

Le risque principal de ce genre d'approche pourrait être celui d'une déperdition de pertinence du fait de la multiplication du nombre de résultats répondant à la requête. Or, et c'est bien là l'intérêt – et la part d'opacité algorithmique – de ce service, Google Universal Search[34] n'apparaît pas conçu comme un service de fusion unidirectionnelle entre bases de données, permettant d'obtenir systématiquement pour chaque requête, un résultat de chaque base interrogée ou interrogeable. De fait, Google insère depuis déjà assez longtemps des résultats de ses bases « livres », « actualités » ou encore « local » tout en haut de certains résultats web. Avec Universal Search, l'objectif est à terme de rendre tout cela transparent pour l'utilisateur, de ne disposer au final que d'un seul calcul de PageRank « universel », tournant sur l'ensemble des bases Google[35]. Alors que nombre d'autres moteurs mettent en

---

[33] http://www.infoworld.com/article/05/11/23/48OPstrategic_1.html Jon Udell.

[34] Communiqué de presse annonçant le lancement du service et détaillant son ambition : http://www.google.com/intl/en/press/pressrel/universalsearch_20070516.html

[35] Des bases de plus en plus nombreuses et conséquentes suite aux rachats effectués. En se rendant par exemple propriétaire du service de partage de vidéos YouTube, Google peut mettre en place une indexation plus rapide et plus efficace de cet extraordinaire fonds documentaire. Dans le même ordre d'idée, rappelons que Yahoo ! est de son côté propriétaire du service de partage de photos FlickR et du site de partage de signets Del.icio.us.

place la même approche[36], la condition *sine qua non* de sa réussite est le passage au premier plan de la gestion de l'historique des recherches individuelles : la pertinence et la hiérarchisation d'un ensemble de contenus hétérogènes n'a de sens qu'au regard des intérêts exprimés par chacun dans le cadre de ses recherches précédentes. Ainsi tel internaute se connectant fréquemment sur des sites de librairies en ligne pourrait voir augmenter significativement le nombre d'items de la base « livre » sur sa page de résultats. La question de la personnalisation, du profilage qu'elle autorise et de l'anticipation qu'elle permet de mettre en place pour tout requête « identifiée » est au cœur de la stratégie de déploiement de l'offre de service des moteurs.

L'autre enjeu d'une telle approche est de rendre l'offre de contenus plus visible grâce à cette verticalisation. Plus précisément il s'agit de porter à visibilité égale des contenus jusqu'ici sous-utilisés ou sous-exploités, pour augmenter leur potentiel marchand en dopant de la sorte le rendement des liens publicitaires afférents.

Ce qui peut être vu comme l'aboutissement logique du phénomène de dérive des continents documentaires a pour effet collatéral de relancer la course technologique autour de l'algorithmie des moteurs, algorithmies qui semblaient arrivées à maturité avec un effet de convergence observable[37]. Non pas qu'il ne soit pas possible des les optimiser davantage[38], mais les coûts en termes d'investissements et de recherche et développement deviennent très élevés au regard du mince gain qualitatif attendu. La recherche universelle ouvre un immense champ de possibles pour la mise en œuvre d'algorithmes capables de prendre en charge les paramètres excessivement complexes de la personnalisation, de la gestion des historiques de recherche, de l'aspect relationnel ou affinitaire qui relie un nombre de plus en plus grand d'items, ou encore du brassage de ces gigantesques silos de donnés. Un brassage totalement inédit à cette échelle.

… à l'algorithmie ambiante en général

A la manière de l'informatique « ambiante » qui a vocation à se diluer dans l'environnement au travers d'interfaces prenant la forme d'objets quotidiens, se dessinent les contours d'une algorithmie également ambiante, c'est à dire mettant sous la coupe de la puissance calculatoire des moteurs, la moindre de nos interactions en ligne, le moindre de nos comportements connectés, la plus infime trace de nos plus éphémères conversations. Un exemple parlant (parmi d'autres) est lisible dans l'analyse du brevet déposé par Google pour son service de recherche de blogs[39]. Parmi la liste des critères retenus pour calculer la pertinence d'un blog par rapport à un autre, il indique que : « (…) *les références vers le blog par d'autres sources peuvent également être des indications positives sur la qualité dudit blog. Par exemple,* ***le contenu des emails ou la transcription de conversations (chat)*** *peuvent contenir l'Url du blog. Les emails ou les clavardages (chat) qui incluent des liens vers le blog sont des indicateurs positifs de la qualité dudit blog.* »

Derrière cette algorithmie ambiante on trouve la volonté déterminée d'optimiser encore davantage la marchandisation de toute unité documentaire recensée, quelle que soit sa sphère d'appartenance d'origine (publique, prive, intime), sa nature médiatique propre (image, son, vidéo, page web, chapitre de livre, etc...), sa granularité (un extrait de livre, un billet de blog, un extrait de vidéo …) et son taux de partage sur le réseau (usage personnel

---

[36] Clusty.com ou Live.com. Pour une liste plus complète, voir le billet : http://www.readwriteweb.com/archives/how_alt_search_engines_implemented_universal_search.php

[37] http://aixtal.blogspot.com/2007/11/moteurs-comparaison-google-yahoo.html

[38] (cf point 1.2)

[39] http://blogsearch.google.com. Le brevet est disponible à l'adresse : http://tinyurl.com/35etd4

uniquement, usage partagé entre « proches », usage partagé avec l'ensemble des autres utilisateurs du service).

2.4. Mon nom est personne.

L'universalité voulue des stratégies de recherche et d'accès à l'information nécessite, pour être rapidement opératoire, de s'appuyer sur l'adhésion des utilisateurs, d'abord en accentuant les possibilités de personnalisation. Cette personnalisation peut revêtir trois formes.

La personnalisation « invisible » ou « transparente » désigne principalement la collecte des logs de navigation ainsi que celle des différentes actions menées par l'usager dans le cadre d'une session pour laquelle il s'est auparavant identifié.

La **personnalisation « persistante »** est un effet corrélé de la première : une fois que vous vous êtes identifié dans un service (webmail de Google par exemple), lorsque vous ouvrez une nouvelle fenêtre ou un nouvel onglet de navigation pour aller interroger le moteur de recherche de la même société, vous « emportez avec vous » votre identification, vous vous trouvez automatiquement identifié et donc reconnu pour les recherches que vous effectuerez sur le moteur, ce qui permet ensuite de récupérer ces éléments pour les verser dans votre profil et dans votre historique de navigation, et ce sans que vous en ayez explicitement exprimé le besoin. Cette activation « par défaut » est une clé importante dans la stratégie des moteurs.

Le troisième type est une **personnalisation participative**, qui nécessite l'adhésion, la participation explicite et librement consentie des utilisateurs. Il s'agit alors d'activer volontairement la procédure d'identification pour accéder aux services de personnalisation proposés, ou bien de proposer aux utilisateurs de décrire (à l'aide de mots-clés ou de tags) les ressources qu'ils ont produites ou qu'ils souhaitent partager avec d'autres. Si l'on prend l'exemple de l'indexation collaborative de ressources (cf infra), les moteurs multiplient ainsi les chances de repérage et d'accès à des contenus en jouant à la fois sur les modes de classement les plus fréquents (par pertinence, par date ou par « popularité » - les contenus les plus accédés, les vidéos les plus vues …), ainsi que sur les mots-clés déposés par les utilisateurs eux-mêmes.

A l'image du « Read/Write Web »[40], le cœur technologique que constitue l'indexation, s'ouvre « en écriture » aux usagers. Il est important de comprendre l'historique de cette ouverture pour mieux en cerner les perspectives.

2.5. Indexation marchande et indexation sociale : le thesaurus comme trésor.

Dans le contexte d'une économie de l'accès et de l'attention (SALA, 2004) (DAVE, 2001) totalement préemptée par les moteurs de recherche, tout est mis en œuvre pour accroître les possibilités de recouper systématiquement les données ainsi collectées, jusqu'à constituer une « base de donnée des intentions » (BATT, 2003) couvrant l'ensemble des données, informations et connaissances indexables.

S'il est vrai que notre monde est et a toujours été documenté, le cœur de cette représentation est l'indexation, humaine ou machinique, qui demeura pendant des siècles hors de toute considération marchande. Avec l'arrivée des liens sponsorisés fonctionnant aussi bien en production (« j'achète un mot ») qu'en réception (« j'affiche une publicité »), il n'est plus un secteur des industriels culturelles qui n'échappe à cette nouvelle dimension

[40] cf le point 2.2

marchande de la représentation et de l'accès à l'information. L'arrivée de nouvelles procédures d'indexation sociale ou « folksonomies »[41] (DEUF, 2006) (ERTZ, 2006) pourrait être lue comme une alternative possible à la situation monopolistique décrite jusqu'ici. Mais les outils qu'elle nécessite sont la propriété des mêmes acteurs. Dans le même temps, ces pratiques achèvent de faire de chacun d'entre nous les médiateurs-indexeurs permanents de toute trace documentaire existante, si infime soit-elle[42].

Indexeurs sans le savoir.

Dans l'article scientifique où ils exposent pour la première fois le principe du PageRank (BRIN, 1998b), les fondateurs de Google indiquent qu'en sus du mode de comptabilité des liens entrants, chaque internaute producteur de contenu agit comme un indexeur à chaque fois qu'il décide de créer un lien hypertexte. Il s'agit certes là d'une sorte de « degré zéro » de l'indexation, mais l'échelle à laquelle il est pratiqué depuis l'invention du web en fait très concrètement le premier allié des moteurs pour optimiser leurs procédures.

L'échec des balises <META>

Profitant de la manne potentielle que représentent ces ressources humaines planétaires d'indexeurs, la possibilité se fait jour de rationaliser et d'homogénéiser un peu l'indexation en proposant – toujours aux mêmes producteurs de contenus – de prendre la main sur l'indexation de leurs sites grâce aux balises <META> déposées dans l'en-tête HTML de la page. C'est l'aspect **communauté de pratique** qui est ici privilégié : on ne parle à l'époque pas encore de Web 2.0 et la production et la mise en ligne de contenus nécessite un minimum de connaissances techniques et informatiques. Mais ces balises <META> s'avèrent rapidement sous-utilisées (faute de maîtriser les compétences techniques nécessaires) et dévoyées dans des logiques de Spamdexing[43]. La majorité des moteurs fera alors le choix d'abandonner complètement leur prise en compte ou de ne s'en servir qu'à la marge[44].

Standardisation et communauté métier.

Ces balises auront cependant permis de jeter les bases d'une standardisation possible et nécessaire, à la condition que cette dernière soit pilotée par une communauté métier. Or depuis l'avènement d'Internet et l'arrivée massive sur le réseau d'informations relevant des sciences et techniques, la question du classement, de la préservation et de l'archivage documentaire n'appartiennent – heureusement – plus aux seuls moteurs de recherche. Le monde des bibliothèques et de la documentation a mis en œuvre des méthodologies (métadonnées), des standards (Dublin Core[45]), et des spécifications permettant d'appliquer à ces contenus un ensemble de principes d'indexation contrôlée.

Folksonomies : le retour de la communauté comme indexeur.

Les folksonomies désignent « *un processus de classification collaborative par des mots-clés librement choisis, ou le résultat de cette classification* »[46]. Elles puisent leur origine

---

[41] Les folksonomies désignent un système de classification collaborative, à l'aide de mots-clés librement choisis.
[42] L'indexation sociale permet de déposer des tags (mots-clés) sur des micro-contenus (billets de blogs, images) ou sur des macro-contenus (sites web ou ouvrages en ligne).
[43] définition Spamdexing : consistant, par exemple, à positionner un site sur de faux mots-clés ou sur les mots-clés d'un concurrent.
[44] La balise <META Description> est ainsi fréquemment utilisée pour afficher le court texte de description qui figure sous le titre d'un site dans une page standard de résultats de recherche.
[45] http://dublincore.org/
[46] Définition extraite du site Wikipedia (http://www.wikipedia.org)

dans le croisement de deux phénomènes renvoyant à des techniques de recherche et de partage de documents. Techniques de recherche et de filtrage tout d'abord. Grâce à des services comme Del.icio.us[47], de nouvelles plateformes d'échange de signets (« social bookmarking ») voient le jour, sur lesquelles chaque utilisateur peut déposer ses signets et les assortir de mots-clés. Il est ensuite possible de laisser un accès et une visibilité complète aux listes constituées pour les partager avec d'autres. Le succès des Folksonomies est conforté par le fait qu'elles s'appuient sur des communautés d'intérêt et d'usage.

Ontologies et Web sémantique

La dernière étape de ce rapide panorama de la problématique de l'indexation sur le Web, est celle du Web sémantique. La communauté visée est ici celle réunissant les compétences, les outils et l'approche métier nécessaire au déploiement d'ontologies au formalisme très rigoureux. Mais ce web sémantique qui apparaissait il y a encore quelques temps comme un simple rêve ou comme un inaccessible idéal, dispose aujourd'hui d'éléments contextuels favorables à son déploiement. Parmi ceux-là : la structuration de plus en plus forte de certains contenus web (cf point 2.2), l'unification des différents gisements informationnels et les options de personnalisation de plus en plus fine qu'elle autorise, mais également le besoin de plus en plus fortement exprimé par les usagers de pouvoir disposer de fonctionnalités de recherche « intelligente », lesquelles ne peuvent être imaginées sans que soit mis en place un formalisme ontologique minimal.

3. Si loin … si proche. Rêves et réalités motorisés.

3.1 Rêve de Web Operating System

Du strict point de vue des usages, le Web semble aujourd'hui n'exister que « dans » et « par » l'image que les moteurs en donnent. A l'évidence, les usagers ne sont plus simplement utilisateurs mais bel et bien acteurs de la production et – dans une moindre mesure – de l'organisation des contenus sur le Net. Ces mêmes usagers bénéficient, dans le cadre de l'offre des moteurs de recherche, d'une gamme d'outils et d'applications de plus en plus riches, et de possibilités d'interactions de plus en plus denses et personnalisables avec des contenus de tout type et de granularité différentes. Le résultat, constaté par les analystes en même temps qu'encouragé par les moteurs, est que nos comportements informationnels basculent chaque jour un peu plus « en ligne ».

Le lieu dans lequel convergent la plupart des applications informatiques que nous utilisons quotidiennement n'est plus nécessairement le système d'exploitation (Operating System) de notre ordinateur personnel, mais bel et bien le web, ou plus précisément l'espace d'accès au web que les moteurs de recherche mettent à notre disposition. C'est bien à la constitution d'un Web Operating System que nous assistons aujourd'hui, et qui peut être caractérisé par la place qu'il laisse au partage (contrôlé) et au mixage d'applications (libres ou propriétaires).

Webtop

Comme nous l'avons déjà indiqué, le modèle économique des moteurs de recherche en particulier et du secteur de l'informatique logicielle grand public en général a changé. Il devient tout aussi voire plus intéressant de mettre à disposition gratuitement des applications

[47] http://del.icio.us

plutôt que de vendre des licences logicielles[48]. Le modèle de la publicité contextuelle et ciblée devient d'autant plus intéressant que l'on attire (par la gratuité des services offerts) un nombre conséquent d'utilisateurs ou de simples visiteurs. Ce « simple » paramètre économique suffit à expliquer les offres de plus en plus alléchantes des moteurs de recherche qui en viennent tous progressivement à proposer des solutions de stockage illimitées (pour nos courriers électroniques par exemple).

Sans engager pour autant une disparition des disques durs (ERTZ, 2005b), il n'y a aujourd'hui plus de barrière entre le local (le disque dur de « ma » machine) et le global (Internet). Les ordinateurs personnels sont depuis déjà longtemps « sur » internet. La nouveauté c'est que leurs contenus sont aujourd'hui « dans » le web, et plus exactement, « dans » les moteurs de recherche. D'où l'anglicisme de « webtop » forgé sur le modèle du « laptop » (ordinateur portable). Les ordinateurs étaient devenus portables. Le web est déjà ubiquitaire.

Collectif

Le passage massif vers des pratiques communautaires de travail se cristallise dans les différents sites et réseaux sociaux/communautaires[49] « motorisés », c'est à dire soit équipés d'une technologie de recherche appartenant à l'un des trois grands opérateurs, soit rachetés par l'un des trois mêmes.

Mixage

Le mixage ne porte pas seulement sur les différents silos informationnels que permet de rassembler la recherche universelle. Il concerne également la combinaison d'applications au travers du phénomène des Mashups, des applications disponibles sur des sites web dont le contenu résulte de la combinaison de plusieurs autres sources ou applications. L'une des plus célèbres permet par exemple de s'appuyer sur le service de cartographie proposé par Google Maps pour visualiser le catalogue d'une agence immobilière.

3.2. Rêve d'implicite

L'importance du chemin.

Un moteur de recherche peut-il anticiper nos requêtes ? La recherche d'information peut-elle revêtir une dimension implicite, c'est à dire s'effectuer sur un besoin non encore exprimé ou formalisé ? Telles sont les questions sur lesquelles travaillent actuellement les moteurs. En observant aujourd'hui le développement des usages et des applications web, on observe une dynamique très forte : les processus et leurs applications « descendent » au niveau de l'usager en s'efforçant, notamment via une personnalisation de plus en plus fine, de rendre parfaitement intuitive la définition de la requête et la ou les stratégies de recherche associées. Les moteurs de recherche ne fonctionnent plus sur un modèle « donne-moi ce que je tape » (travail sur l'occurrence des mots-clés choisis) mais « donne-moi ce que je veux » (travail sur l'adéquation des résultats de recherche au profil de l'usager, ou au profil d'un macro-ensemble de requêtes semblables).

Il s'agit de transformer en itinéraire dirigé et centré sur les attentes de l'usager, ce qui était considéré au début du web comme une nuisance, à savoir la profusion de parcours

[48] sauf naturellement à se retrouver en situation de monpole ou de quasi-monopole
[49] MySpace (http://myspace.com) et Facebook (http://facebook.com) en sont emblématiques

autorisée par les liens hypertextes[50]. Qui aurait pu imaginer il y a encore quelques années qu'une interface de recherche soit capable, sur la base d'une simple requête, de nous fournir en retour non plus de simples « résultats », mais des recommandations, des choix de reformulation, en accord avec nos choix, nos itinéraires, nos parcours précédents ou avec ceux d'une collectivité ayant déposé une requête semblable ? Nous assistons là au retour à l'idée première de l'hypertexte telle qu'elle avait été théorisée par (BUSH, 1945) : le parcours, le chemin (« trail ») importe au moins autant que l'instanciation du lien. Nous sommes donc passés d'une toute puissance du lien hypertexte, point nécessairement nodal de développement du réseau et des services et outils associés, à une toute puissance du « parcours », de la navigation « qui fait sens », de la navigation « orientée » au double sens du terme. Et les premiers résultats sont là : au moment où, sur Amazon par exemple, nous « activons » les liens proposés sous forme de recommandation[51] suite à une requête ou une recherche initiale, nous n'avons pas formulé explicitement ce besoin. Au final pourtant, le parcours « aura fait sens » (avec plus ou moins de succès), et l'activité mentale couplant recherche et navigation n'aura plus eu besoin d'être littéralement « déclarative », permettant ainsi de parler d'un web implicite.

Myware + everyware

Demain probablement, ces mêmes interfaces, ces mêmes moteurs, sauront ce que nous sommes le plus susceptibles de chercher ou de saisir comme requête selon l'heure de la journée, le lieu de notre connexion ou encore nos historiques de recherche, et ce sans même avoir besoin d'une requête initiale, d'un premier « amorçage ».

La langue anglaise étant en la matière plus synthétique et illustrative que la nôtre, on pourrait décrire ce futur (proche) par la combinaison de deux termes : Myware + Everyware. « Myware » pour ce cortex collectif, in-vivo. « everyware » pour désigner le règne de cette algorithmique ambiante, ubiquitaire.

3.3. Rêve sémantique

La question d'un web sémantique et de moteurs susceptibles d'en extraire plus « intelligemment » du contenu, ne cesse de se poser depuis l'article fondateur de (BERN, 2001).

Moteurs sémantiques : l'approche « top-down »

Les moteurs de recherche « sémantiques » soulèvent plusieurs questions. La première est celle de la relative complexité de leur prise en main à l'heure où les utilisateurs réclament des interfaces de plus en plus fluides, riches[52] et intuitives. On sait par exemple le rôle que joua la sobriété et la simplicité de l'interface de Google dans son succès. On connaît également la difficulté que posent les interfaces cartographiques[53] à l'internaute lambda. Même les outils de catégorisation[54] ont mis longtemps avant d'être adoptés par le grand public et sont encore aujourd'hui perçus comme plutôt réservés à des utilisateurs avertis. Si tant est que le web sémantique puisse un jour être réalisé dans la forme imaginée par son concepteur, encore faudra-t-il que les moteurs proposent des interfaces adaptées.

---

[50] c'est le « lost in hyperspace problem » défini par (CONK, 1987) et reliant la notion de « navigation » à celle de « désorientation » et de surcharge cognitive.

[51] Sur le mode « les gens ayant acheté ce disque ont aussi achetés ceux-là »

[52] sur la notion d'interface riche ou RIA (Rich Internet Application), voir http://fr.wikipedia.org/wiki/Rich_Internet_Application

[53] Kartoo (http://www.kartoo.com), Grokker (http://www.grokker.com)

[54] Exalead (http://www.exalead.com), Clusty (http://www.clusty.com)

La richesse d'un web sémantique, du point de vue de la recherche d'information, se situe principalement dans les capacités de navigation optimisées qu'il permettrait d'offrir. S'il s'agit « simplement » de répondre à des questions du type « Quid »[55] ou même de simple désambiguïsation[56], les moteurs de recherche actuels gèrent suffisamment bien ce genre de questions. En revanche à partir d'une requête initiale, le fait de pouvoir naviguer non plus simplement à l'aveugle ou sur la base des backlinks menant d'un site à un autre, mais bel et bien dans un environnement sémantique explicite et contextualisé pourrait être grandement intéressant[57].

Si une réelle recherche sémantique devient un jour possible, elle sera longtemps réservée à l'exploration de corpus dédiés dans des contextes de tâche bien identifiés et au sein de communautés de pratique très délimitées, avant de se trouver à portée d'interface du grand public.

L'état de l'art actuel indique plutôt que les avancées technologiques se servent du web pour proposer une architecture de navigation inspirée de celle des bases de données relationnelles.

Moteurs sémantiques : l'approche « bottom-up ».

A l'inverse d'une approche descendante impliquant que soient déjà franchis les différents obstacles techniques permettant la mise en œuvre d'un web totalement sémantique, l'évolution des fonctionnalités sémantiques des moteurs de recherche suivra plus probablement une approche ascendante, émergente. Il s'agit cette fois de prendre progressivement en compte les différentes avancées des protocoles, langages et formalismes liés au web sémantique, non pas de manière globale mais sur des contenus très ciblés, ou dans le cadre de contextes de recherche là encore très spécialisés.

En Mars 2008, Yahoo ! a ainsi annoncé[58] qu'il prendrait en compte le standard RDF[59] ainsi que les microformats[60]. Pour ne prendre que ce dernier exemple, de nombreux développements existent actuellement[61]. La dernière course de fond engagée par les moteurs consistera donc à en prendre le maximum en compte (sans nécessairement attendre une harmonisation globale ou une standardisation univoque de l'ensemble des développements applicatifs en cours), tout en trouvant le moyen de s'en servir pour « enrichir » l'expérience utilisateur lors d'une recherche d'information, par exemple en présentant des résultats de recherche davantage structurés ou permettant davantage d'interactions synchrones avec d'autres recherches, d'autres services, d'autres terminaux d'accès[62]. Dit autrement, les moteurs sémantiques pourraient fournir une solution aux limitations de la recherche par mot-clé.

---

[55] date de naissance de telle célébrité, date de telle guerre ou catastrophe naturelle, définition de tel terme …

[56] sur des requêtes de type « jaguar » nécessitant de préciser s'il s'agit de l'animal ou du véhicule

[57] Le billet d'Alex Iskold sur les mythes et la réalité de la recherche sémantique est sur ce point particulièrement éclairant : http://alexiskold.wordpress.com/2008/05/30/semantic-search-the-myth-and-reality/

[58] http://www.techcrunch.com/2008/03/13/yahoo-embraces-the-semantic-web-expect-the-web-to-organize-itself-in-a-hurry/

[59] « Resource Description Framework (RDF) est un modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs métadonnées, de façon à permettre le traitement automatique de telles descriptions. Développé par le W3C, RDF est le langage de base du Web sémantique. » Source : http://fr.wikipedia.org/wiki/Resource_Description_Framework

[60] Les microformats sont un langage de balisage qui permet l'expression de la sémantique dans une page web HTML (ou XHTML).

[61] L'un des plus prometteurs est FOAF (Friend of A Friend), un vocabulaire RDF permettant de décrire des personnes et les relations qu'elles entretiennent entre elles. (http://fr.wikipedia.org/wiki/FOAF).

[62] Téléphones cellulaires notamment

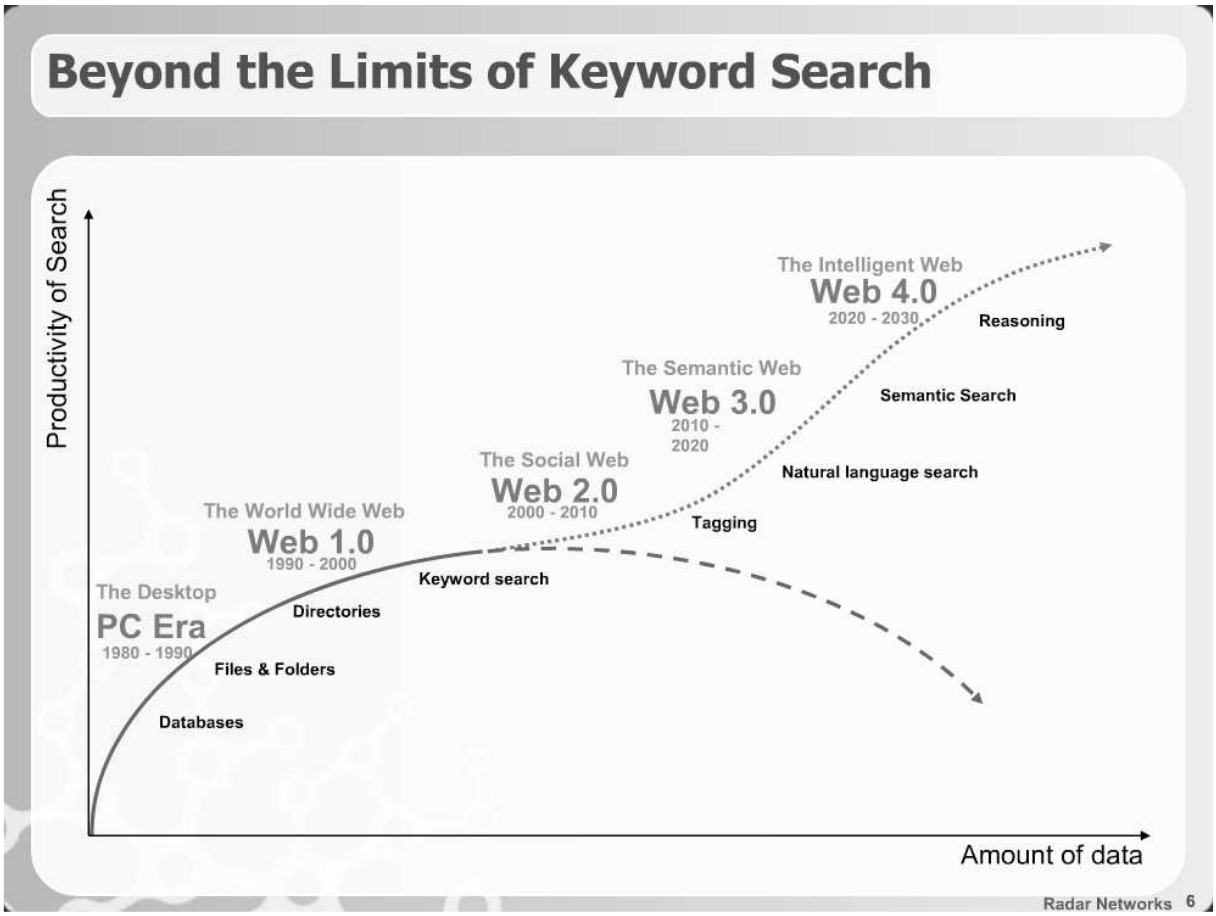

Figure 1 : Les limites de la recherche par mot-clé.[63]

Petite revue de troupes

Il existe actuellement trois grands types de moteurs prenant en charge certains aspects du web sémantique.

Les moteurs généralistes et grand public, qui intègrent progressivement différents microformats.

Les moteurs « sémantisés » dont la plupart n'interrogent pour l'instant qu'une seule base, celle de Wikipedia, et dans lesquels les requêtes s'effectuent sous forme de questions en langage naturel. Une contextualisation et des compléments sémantiques de chaque requête sont alors proposés aux utilisateurs. Dans cette catégorie citons Trueknowledge (http://www.trueknowledge.com/), Powerset (http://www/powerset.com) ou encore Twine (http ://www.twine.com/). Ces moteurs sont encore en phase expérimentale (versions alpha ou béta). La question clé est celle du niveau de structuration de leur base d'appui. La proposition de fonctionnalités sémantiques nécessite en effet, en plus d'une algorithmie particulière, des corpus relativement structurés ou aisément « structurables ».

Il faut enfin citer, pour être complet, les moteurs de recherche et solutions « corporate » qui intègrent des fonctionnalités sémantiques souvent plus riches car ils

[63] Image extraite du diaporama de Nova Spivack, http://novaspivack.typepad.com/nova_spivacks_weblog/files/nova_spivack_semantic_web_talk.ppt

fonctionnent généralement sur des silos informationnels bien délimités (les données internes de l'entreprise).[64]

Au final, ce « rêve sémantique » se donne aujourd'hui à lire dans une réalité de plus en plus « sémantisée » de la recherche d'information.

3.4. Une question de génération ...

Les moteurs de première génération, reposaient sur l'analyse du contenu des pages indexées et ne prenaient en compte que le « matching » (croisement) entre des mots-clés (ceux de la requête et ceux des pages indexées).

Les moteurs de deuxième génération s'intéressent d'abord à la structure du web en tant que graphe orienté, structure qui vînt trouver sa place au cœur de leur algorithmie (PageRank), permettant une recherche d'information « augmentée ».

La troisième génération est celle des moteurs « sociaux » et « personnalisables »[65] qui ajoutent le filtrage amont des sources par les usagers comme un modèle de pertinence différent. La personnalisation permettant à chacun de fabriquer « son » propre moteur[66], de choisir « ses » sources et d'ouvrir ensuite cet espace à d'autres (communautarisation) est aujourd'hui une tendance très forte qui fait suite à l'engouement pour les sites de signets partagés dans lesquels l'usager construit, agrège ses sources et les fait partager à une communauté.

La quatrième génération s'inscrit dans une logique disruptive avec l'arrivée des Mashups[67] et le mixage qui peut opérer entre services, ou entre un moteur et un/des services. Une arrivée qui vient se conjuguer à la structuration de plus en plus fine des contenus, que cette structuration soit produite en aval par les usagers (indexation sociale), ou en amont par les moteurs (microformats et sémantisation des logiques et standards de description – triplets RDF notamment). Par ailleurs, en mettant à disposition un nombre de plus en plus grand d'APIs[68], les moteurs deviennent les premiers fournisseurs de services ... et les premiers bénéficiaires des données collectées.

A la fois cause et conséquence de cette évolution, les techniques liées au protocole RSS viennent offrir au périmètre de la recherche d'information une granularité quantitative et qualitative jusqu'ici impossible à atteindre (ERTZ, 2005c). Les moteurs de recherche reprenant là encore à leur compte ces nouvelles logiques d'exploitation en proposant certains de leurs résultats directement au format RSS (ERTZ, 2008d pp. 30-42).

3.5. Rêve de synchronicité

La dynamique du web, l'émergence de ses contenus s'est très longtemps contentée de se décliner dans un « présent antérieur » de la recherche d'information, les moteurs s'efforçant d'indexer le plus rapidement possible l'essentiel des contenus disponibles, effaçant au fur et à mesure les contenus les plus anciens de leurs index pour les remplacer par des nouveaux.

---

[64] pour un panorama plus complet de cette offre, voir le dossier du Monde Informatique sur le sujet : http://www.lemondeinformatique.fr/dossiers/lire-le-web-semantique-un-vaste-terrain-d-applications-encore-a-defricher-25-page-3.html

[65] cf dans le point 1.1 ceux cités dans la partie « Prescription ».

[66] voir la note 9

[67] Le terme de « Mashup » sert à désigner le couplage de deux applications informatiques pour en construire une troisième.

[68] les API ou Application Program Interface (Interface pour langages de programmation), permettent à une application d'accéder à des programmes pour communiquer ou extraire des données. « Techniquement, les API consistent en des appels de fonction, accompagnés des paramètres à communiquer à l'extérieur de l'applicatif. » (Source : http://www.journaldunet.com/encyclopedie/) Une liste non exhaustive des API de Google est accessible sur http://www.webrankinfo.com/actualites/200601-liste-api-google.htm.

Or le web, les moteurs et leurs usages se déclinent aujourd'hui dans le cadre d'une temporalité « enrichie ». Il est en effet possible, grâce à l'invention de la Wayback Machine[69], de remonter dans le temps pour visualiser les différentes versions antérieures de tel ou tel site. Le passé est ainsi rendu accessible. Par ailleurs, comme nous venons de le détailler, les travaux du web sémantique, la personnalisation sans cesse augmentée et la référence à une dimension implicite de la recherche d'information permettent de penser qu'un « futur proche » ou à tout le moins un « futur immédiat » de la recherche d'information sera un jour accessible grâce aux moteurs de recherche qui relèveront ce pari.

Mais dans la quête perpétuelle d'un espace-temps cohérent et dédié à la recherche d'information en particulier et à nos comportements informationnels en général, une chose manque encore : la possibilité de synchroniser nos informations, nos documents, nos données, nos pratiques, à la fois dans le monde connecté et dans le monde non-connecté. Après avoir réussi à rassembler en un seul et même espace l'ensemble des continents documentaires existants, il importe aujourd'hui de pouvoir y accéder de manière synchrone, c'est à dire de permettre aux outils, aux applications et aux moteurs que nous utilisons quotidiennement, de savoir, « où nous en étions restés », pour pouvoir nous proposer de travailler sur « la dernière version » de tel document partagé, sur « le dernier relevé » de courrier électronique, et ainsi de suite. Des applications comme Google Gears[70] s'inscrivent dans cette logique. L'enjeu est de taille : effacer la dernière différence[71], la dernière frontière, celle entre les usages connectés et non-connectés.

CONCLUSION

Implicite, sémantique, sémantisé, synchrone, mixé et remixé (mashups), applicatif, ubiquitaire, granulaire, collaboratif … les adjectifs ne manquent pas pour dresser l'inventaire des progrès actuels et des ambitions futures des outils d'accès et de recherche d'information. Mais l'évolution des moteurs de recherche ne peut être observée ou même pensée en dehors d'une approche plus globale sur la nature des contenus et des usages sur le Net. Elle ne peut pas davantage l'être sans prendre en compte les paramètres économiques qui bouleversent actuellement des pans entiers des industries culturelles. A ce titre, les moteurs de recherche sont de formidables catalyseurs d'un ensemble beaucoup plus vaste de problématiques ; ils font figure d'objets d'étude à part entière et ce bien au-delà de leur simple dimension algorithmique. Ils interrogent, au double sens du terme, des pans entiers de nos vies numériques.

Qu'il s'agisse de vie publique, de vie privée, mais également de l'offre de substitution qu'ils proposent déjà sur des secteurs et des enjeux de premier plan comme la médecine ou la génomique[72], l'urgence est de poser la question de l'opacité des algorithmes conjuguée à celle de leur omniprésence. Il est aujourd'hui absolument nécessaire d'ouvrir un débat autour de l'écosystème non plus simplement documentaire mais politique qu'ils représentent. Faute de

---

[69] Littéralement « machine à remonter le temps », la Wayback Machine est le moteur de recherche de l'Internet Archive (http://www.archive.org), une fondation s'étant donnée pour mission d'archiver le web. La Wayback Machine donne aujourd'hui accès à un fonds de deux milliards de pages « disparues » ou dans des versions antérieures impossibles à retrouver autrement. Par ailleurs, la question d'un archivage systématique et raisonné est une question essentielle dont se sont aujourd'hui saisies les bibliothèques, en définissant notamment la notion d'un dépôt légal pour les sites web (voir les pages de la BnF sur le sujet : http://www.bnf.fr/pages/infopro/depotleg/dl-internet_quest.htm)

[70] Google Gears permet, en mode « déconnecté », d'accéder à des applications et services connectés (Google Reader – lecteur RSS – et Google Docs pour l'instant)

[71] on pourrait même parler de dernière « différance » dans l'optique de Derrida.

[72] A ce sujet, voir (ERTZ 2008c)

quoi, devant des usages de plus en plus conditionnés à des logiques marchandes instrumentalisant les données collectées au cours même du processus de recherche, c'est l'usager et lui seul qui, dans l'ignorance de ces logiques, se trouvera totalement instrumentalisé, devenant dans un mouvement réflexif paradoxal, le seul et unique objet de sa recherche.

Pour ne prendre qu'un seul exemple lié à l'explosion du phénomène des réseaux sociaux qui « ouvrent » désormais l'immense catalogue des individualités humaines qui les composent à l'indexation par les moteurs de recherche, et qui proposent par ailleurs des options de sémantisation déjà efficientes[73], c'est la question de la pertinence des profils humains qui se trouve posée, celle de savoir si l'Homme est ou non, un document comme les autres[74].

Avec leur perpétuel effet miroir, dans le panoptique sans cesse actualisé qu'ils mettent à notre disposition mais dans lequel ne se donnent à lire que leurs propres dispositions, les moteurs font basculer la question de la politique documentaire dans une dimension et à une échelle inédite, celle d'une macro-documentation du monde, aujourd'hui nécessairement politique.

Bibliographie

(Tous les liens actifs au 7 Mai 2008)

- (BATT 2003) Battelle J., « *The Database of Intentions* », in Searchblog, 13 Novembre 2003. <http://battellemedia.com/archives/000063.php>
- (BERG 2001) Bergman M.K., « *The Deep Web : Surfacing Hidden Value* », Ann Arbor, MI : Scholarly Publishing Office, University of Michigan, University Library, vol. 7, no. 1, August, 2001. <http://hdl.handle.net/2027/spo.3336451.0007.104>
- (BERN 1989) Berners Lee, T., « *Hypertext and the CERN* », 1989. <http://www.w3.org/History/1989/proposal.html>
- (BERN 2001) Berners Lee T., Hendler J., Lassila O., « *The Semantic Web* », in Scientific American, Mai 2001, <http://www.sciam.com/article.cfm ?id=the-semantic-web>, traduction française accessible in La Lettre de l'Urfist n°28 <http://www.urfist.cict.fr/archive/lettres/lettre28/lettre28.pdf>
- (BERN 2007) Berners Lee, T., « Giant Global Graph », 21 Novembre 2007. <http://dig.csail.mit.edu/breadcrumbs/node/215>
- (BRIN 1998a) Brin S., Motwani R., Page L., Winograd T., « What can you do with a Web in your pocket ? » Data Engineering Bulletin, n° 21, pp. 37–47, 1998.
- (BRIN 1998b) Brin S., Page L., « The anatomy of a large-scale hypertextual Web search Engine ». Computer Networks and ISDN Systems, vol. 33, pp. 107–117, 1998.
- (BRIN 1999) Brin S., Page L.,, Motwami R., Winograd T., « The PageRank citation ranking: bringing order to the Web. » Technical Report 1999-0120, Computer Science Department, Stanford University, 1999.
- (BROD 2002) Broder, A. 2002. « A taxonomy of web search. » SIGIR Forum 36, 2 (Sep. 2002), 3-10. <http ://doi.acm.org/10.1145/792550.792552.>
- (BUSH 45) Bush V., « *As We May Think.* », pp. 101-108, in The Atlantic Monthly, vol.1, n°176, Juillet 1945. <http://www.isg.sfu.ca/~duchier/misc/vbush>
- (CONK 1897) Conklin J., Hypertext : An introduction and survey. Computer Magazine, 20, 17-41, 1987.

[73] microformat FOAF notamment (cf supra)
[74] (ERTZ, 2008a)

- (CROS 2008) Crosnier, H. le, « Mouvements tectoniques sur la toile », in Le Monde Diplomatique, Mars 2008, p.19. < http ://www.monde-diplomatique.fr/2008/03/LE_CROSNIER/15673>
- (DAVE 2001) Davenport T., Beck J. C., The Attention Economy. Harvard Business School Press, 2001.
- (DAVI 2006) David S., Pinch T., « Six degrees of reputation : The use and abuse of online review and recommendation systems », in First Connaître, vol.11, n° 3 , Mars 2006. http ://firstmonday.org/issues/issue11_3/david/index.html
- (ERTZ 2004) Ertzscheid O., Gallezot G., « *Des machines pour chercher au hazard : moteurs de recherché et recherche d'information.* » XIVe Congrès SFSIC, Questionner l'internationalisation : cultures, acteurs, organisations, machines, Béziers 2004. <http ://archivesic.ccsd.cnrs.fr/sic_00000989/fr/>
- (ERTZ 2005a) Ertzscheid O., « *Google se Microsoftise* » in Affordance.info, 27 Septembre 2005. <http ://affordance.typepad.com/mon_weblog/2005/09/google_se_micro.html>
- (ERTZ 2005b) Ertzscheid, O., *« Le jour où notre disque dur aura disparu.* », in Le Monde, 21 Avril 2005.
- (ERTZ 2005c) Ertzscheid O., « *Weblogs : un nouveau paradigme pour les systèmes d'information et la diffusion de connaissances ? Applications et cas d'usage en contexte de veille et d'intelligence économique.* », Communication avec Actes, Colloque ISKO 2005, Nancy, en ligne : http ://archivesic.ccsd.cnrs.fr/sic_00001433
- (ERTZ 2006) Ertzscheid O., Gallezot G., « *Etude exploratoire des pratiques d'indexation sociale comme une renégociation des espaces documentaires.* » in Document numérique et société, (sous la dir. De) Ghislaine Chartron et Evelyne Broudoux. ADBS Éditions, 2006. 344 p. Collection Sciences et techniques de l'information.
- (ERTZ 2008a) Ertzscheid O., « Bienvenue dans le World Life Web », in Ecrans.fr, 15 février 2008. <http://www.ecrans.fr/Bienvenue-dans-le-World-Life-Web,3016.html>
- (ERTZ 2008b) Ertzscheid O., « *L'industrie de la recommandation est-elle recommandable ?* » Communication aux journées d'étude Polyphonies du livre, Mars 2008. Diaporama. <http ://www.slideshare.net/olivier/industrie-de-la-recommandation/>
- (ERTZ 2008c) Ertzscheid O., « *Docteur Google : quelle médecine pour demain ?* », in Affordance.info, 2 Mars 2008. <http://affordance.typepad.com/mon_weblog/2008/03/docteur-google.html>
- (ERTZ 2008d) Ertzscheid O., Créer, trouver et exploiter les blogs. Paris, ADBS Editions, 2008, 64p. (Coll. L'essentiel sur)
- (GARF 1972) Garfield E, « Citation analysis as a tool in journal evaluation », Science, vol. 178, pp.471-479, 1972.
- (GREE 2005) Green Adam, « *2006 : The Year the Web explodes* », in Darwinian Web, 18 Novembre 2005. < http ://darwinianweb.com/archive/2005/78.html>
- (JING 2008) Jing Y., Baluja S., « PageRank for product image search », in WWW2008, Beijing, 21-25 Avril 2008. < http ://www.esprockets.com/papers/www2008-jing-baluja.pdf>
- (LANG 2003) Langville Amy N., Meyer Carl D., « Deeper Inside PageRank », Internet Mathematics. Vol. 1, no. 3, pp. 335–380, 2003.
- (DEUF 2006) Le Deuff O., « *Folksonomies : Les usagers indexent le web.* », BBF, 2006, n° 4, p. 66-70, <http://bbf.enssib.fr>
- (LYMA 2003) Lyman P., Varian Hal R., « *How Much Information* », 2003. < http ://www.sims.berkeley.edu/how-much-info-2003>.
- (NIEL 2005) Nielsen J., « *One billion Internet Users.* », Décembre 2005. En ligne : http ://www.useit.com/alertbox/internet_growth.html.
- (SALA 2004) Salaün J.-M., « *Libre accès aux ressources scientifiques et place des bibliothèques.* », BBF, 2004, n° 6, p. 20-30, <http://bbf.enssib.fr>.
- (TISS 2001) TISSERON S., L'intimité surexposée, Paris, Editions Ramsay , 2001.
- (VERO 2005) Véronis J., « *Google : index mystère* », in Technologies du langage, 27 Septembre 2005. <http://aixtal.blogspot.com/2005/09/google-index-mystre.html>