Que sais-je ?

L'INFORMATION ET LE RENSEIGNEMENT PAR INTERNET

Laurence Ifrah

puf

QUE SAIS-JE ?

L'information et le renseignement par Internet

LAURENCE IFRAH

Table des matières

L'information et le renseignement par Internet
Chapitre I
Le renseignement public et privé
I. L’intelligence économique
II. Les technologies de l’information
III. Le Web 2.0
IV. Del.icio.us ou le « social-bookmarking »*
1. Twitter
2. Le Web, c’est aussi l’ère du mouvement permanent, d’une versatilité très rapide des usages
3. Plus de bruit que de mots
V. L’opinion contre l’expertise
VI. Le renseignement étatique par le Web 2.0
Notes
Chapitre II
Histoire des moteurs de recherche
I. Intégration du Web 2.0 dans les moteurs de recherche
II. Étude comparative
1. 1re recherche
2. 2e recherche
Notes
Chapitre III
La veille
I. Les différents types de veille
1. La veille juridique
2. La veille concurrentielle
3. La veille technologique
4. La veille marketing
5. La veille stratégique
II. L’e-réputation
III. L’e-réputation des États ou la corruption du buzz numérique
Notes
Chapitre IV
L’accès à, et la manipulation de l’information
I. Les manipulations boursières, les fausses rumeurs
II. L’espionnage industriel
III. Les dérives criminelles (hackers et mafias)
Notes
Chapitre V
Le Web 2.0 et les réseaux sociaux, l’envers du décor
I. Collecte d’informations militaires
II. Facebook, MySpace, une source pour les services de renseignements comme pour les malfaiteurs
III. Portrait d’un inconnu : la fin de la vie privée
IV. Usurpation d’identité

V. Les réseaux sociaux et l’entreprise
Notes
Chapitre VI
Le contrôle de la validité de l’information
I. L’absence de confidentialité des recherches sur Internet
II. Recoupement des requêtes
III. Les cookies
IV. Anonymat et préservation de la vie privée
Notes
Chapitre VII
Organisation de l’information
I. L’information structurée
II. L’information non structurée
III. Les applications de type Desktop Search
1. Google Desktop Search
2. Copernic Desktop Search
3. Windows Desktop Search
4. Exalead Desktop Search
IV. Comparatif
V. Les risques liés aux outils de recherche Desktop
VI. La recherche d’informations dans une organisation
Notes
Chapitre VIII
Un savoir-faire français
I. Numérisation
II. Recherche sémantique
1. Recherche « intelligente » d’informations
2. Exalead
III. Text mining
1. Temis
2. Arisem
IV. Veille
1. Digimind
2. Ami Software
3. kb Crawl
V. Linguistique : l’analyse sémantique verticale
Chapitre IX
Perspectives
I. En entreprise
II. Dans le monde du Search
Notes
Conclusion – Dépendance : le risque d'une rupture
Notes
Glossaire
Bibliographie
Bibliographie et Webographie

Chapitre I

Le renseignement public et privé

Le renseignement est à la fois une définition relative à des informations intéressant une structure publique ou privée pour la formulation et l'application de ses politiques de sécurité. Le renseignement permet de préserver ses intérêts, et de gérer les menaces provenant d'adversaires avérés ou potentiels [1]. C'est aussi une organisation visant à la collecte et à l'analyse de ces informations.

En France, c'est sous le règne de Louis XIII (1601-1643) qu'a été constitué le premier service de renseignements français moderne, dirigé par Richelieu. Sous Louis XV (1710-1774), se crée le Secret du roi, dont le chef, le comte Charles de Broglie, organise un puissant réseau de renseignements et d'interception du courrier. Après la chute de l'Ancien Régime (sous le Consulat et le Premier Empire), Fouché (1759-1820), ministre de la Police, et Talleyrand (1754-1835), ministre des Relations extérieures, furent des personnages clés du renseignement français sous les ordres de Napoléon qui attache une grande importance à l'information secrète.

Les agences privées de renseignements et de recherche existaient dès le XIX^{e} siècle (avec notamment le fameux Vidocq en matière financière). Leurs missions concernaient autant l'adultère que la Bourse.

Née à Sedan lors de la défaite de 1870, la section de statistiques et de reconnaissances militaires (service de contre-espionnage) a été fondée par décret, le 8 juin 1871, et placée sous la tutelle du 2^{e} Bureau de l'état-major [2]. Le renseignement français s'est considérablement développé au début des années 1900 et a démontré ses capacités en matière de collecte d'information et dans le décryptage des transmissions chiffrées lors de la Première Guerre mondiale. En 1943 naît la dgss (Direction générale des services spéciaux), remplacée l'année suivante par la dger (Direction générale des études et recherches). C'est à la fin de l'année 1945 qu'est créé le sdece (Service de documentation extérieure et de contre-espionnage) qui deviendra la dgse (Direction générale de la sécurité extérieure) en avril 1982.

Aux États-Unis, c'est le 5 juillet 1865, à Washington dc, qu'a été créée la Secret Service Division pour lutter, à l'origine, contre la fausse monnaie. Une mission étendue dès 1867 à la fraude en général ou à toute personne risquant de porter atteinte au gouvernement américain [3].

Au cours de la guerre froide, les activités des services secrets se sont considérablement développées, mettant face à face l'Est et l'Ouest. À la fin de ce conflit, certains de ces services ont été réorientés vers le renseignement consacré à la défense des entreprises privées [4] et ont transposé leurs méthodes dans le domaine économique.

Au début du XX^{e} siècle, les services américains et anglais ont démontré leurs compétences technologiques en en mettant sur écoutes les câbles téléphoniques sous-marins installés à partir de 1850 [5].

La nsa (National Security Agency), fondée en mai 1949, est spécialisée dans l'interception des

communications privées et publiques (sigint). Cette agence est responsable de la collecte et de l'analyse de toutes formes de communications, aussi bien gouvernementales que commerciales ou encore personnelles, par tout mode de transmission. Elle est à l'origine du système d'espionnage des communications, le réseau Échelon, qu'elle gère avec les services de renseignements des États membres de l'ukusa (United Kingdom usa Security Agreement), le Canada, l'Australie et la Nouvelle-Zélande. Toutes les informations collectées sont ensuite analysées au quartier général de la nsa à Fort George G. Meade dans le Maryland, aux États-Unis.

I. L'intelligence économique

Ce sont les Japonais qui, les premiers, ont lancé le concept d'intelligence économique au début des années 1950. Le miti (Ministry of International Trade and Industry) et le jetro (Japan External Trade Organisation) avaient pour objectif de les aider à relancer l'économie quelques années après la fin de la Seconde Guerre mondiale. Les Américains s'y sont intéressés dans les années 1980 avec les travaux de Michael Porter, professeur à Harvard en stratégie d'entreprise. Les deux points marquants de ses recherches ont été la prise de conscience d'un marché désormais mondial et non plus national ainsi que l'importance des moyens techniques nécessaires à la collecte et à l'analyse d'informations pour les organisations. En France, où, pendant longtemps, le terme d'intelligence économique a donné lieu à quelques mauvaises interprétations, la prise de conscience a été plus longue : les premières tentatives ont été initiées à la suite de la publication du rapport Martre en 1994. En 1995, le Premier ministre, Édouard Balladur, crée le Comité pour la compétitivité et la sécurité économique. En 2003, le gouvernement français engage une réelle politique publique en matière d'intelligence économique déclinée à partir des propositions du rapport du député Bernard Carayon. Alain Juillet est nommé, le 31 décembre 2003, par décret du président de la République, haut responsable en charge de l'intelligence économique auprès du secrétaire général de la Défense nationale [6].

Les agences privées de renseignements et de recherche existaient dès le XIX^{e} siècle [7]. La plupart ont évolué vers le renseignement pour les entreprises, ou l'intelligence économique. Bien que soumise à des règles strictes d'éthique et de déontologie, l'information sur les concurrents, leurs projets, leurs produits et leurs salariés a souvent été l'objet de dérives de la part d'officines privées. Ces dernières se heurtent à la concurrence que certains qualifient de déloyales des agences de renseignements étatiques. Il s'agit dès lors de définir comment mettre au service des entreprises les connaissances obtenues par l'administration sans atteinte à la libre concurrence des acteurs du privé.

II. Les technologies de l'information

Au début des années 1990, l'accès à Internet s'est démocratisé, les foyers ont commencé à s'équiper d'ordinateurs et à installer une connexion vers « le réseau des réseaux ». Le débit était si faible que peu d'entreprises se risquaient à mettre en ligne un site présentant quelques pages d'information dont le temps d'affichage particulièrement long suffisait à décourager le plus motivé des internautes. Ce qu'il y avait de plus attrayant était surtout les forums. Les internautes passaient des heures en ligne à échanger avec des correspondants du monde entier sur les sujets les plus divers. À partir de 1997, des informations souvent inédites apparaissent sur le Net qui devient alors une source de renseignements. Cependant, le nombre d'internautes est assez faible en France et atteint à peine 4 millions en 1998, du fait du Minitel, encore très présent dans les foyers et les entreprises. En 2004, le nombre de connectés est de 13,78 millions en France et 294,48 dans le monde [8]. La notion de recherche d'information sur la Toile n'est pas encore

acquise.

À présent, le cap symbolique de un milliard d'internautes a été franchi (en janv. 2009 [9]), naviguant sur environ un billion de pages Web. Désormais, entreprendre une recherche d'information sur Internet est un geste quasi quotidien, un réflexe naturel, mais le foisonnement des données est tel qu'il devient beaucoup plus complexe, pour les internautes, de trier le bon grain de l'ivraie.

III. Le Web 2.0

Le Web 2.0 est un espace collaboratif relationnel, un réseau d'interaction sociale dont la définition est multiple parce qu'il a de multiples dimensions. Qualifié parfois de buzzword, il est en fait un nouveau modèle rédactionnel qui tire parti de l'intelligence collective. Wikipédia, Twitter, MySpace, Del.icio.us ou encore Flickr en sont les meilleurs exemples.

Le concept de Web 2.0 est né lors d'une conférence de brainstorming organisée par Dale Dougherty (de la société d'édition O'Reilly) et Craig Cline (de Mediative International), destinée à développer des idées pour marquer l'émergence d'une nouvelle étape dans la courte histoire du Web caractérisée par un changement des règles et une modification des modèles de revenus.

L'implication des utilisateurs devient un facteur clé qui s'est développé lors de la création de la blogosphère dont l'apparition a bouleversé Internet dans ses fondements. La croissance d'un site en termes de popularité devient relative au nombre de ses participants qui peuvent être soit uniquement consommateurs, soit contribuer à l'amélioration du service offert. Mélange hétéroclite de technique et de social, le Web 2.0 replace l'utilisateur et ses réseaux sociaux au centre d'internet, passant ainsi de la notion de produit à celle de service. « Le Web 2.0 repose sur un ensemble de modèles de conception : des systèmes architecturaux plus intelligents qui permettent aux gens de les utiliser, des modèles d'affaires légers qui rendent possible la syndication* [10] et la coopération des données et des services… Le Web 2.0 c'est le moment où les gens réalisent que ce n'est pas le logiciel qui fait le Web, mais les services [11] ! »

Ce tableau illustre la transition qui s'est opérée entre le Web 1.0 et le 2.0. Les blogs, par exemple, sont la démonstration idéale de cette mutation. Les sites perso du Web 1.0 – des pages statiques publiées par des utilisateurs – sont devenus des forums d'échange sur lesquels les internautes réagissent au centre d'intérêt du blogueur, en apportant leur contribution sous la forme de billets ou de commentaires.

Les nombreux liens habituellement présents sur les pages étayent les propos de l'auteur en orientant ses lecteurs vers d'autres blogs traitant du même sujet, créant ainsi un maillage interconnecté vers des domaines d'expertises.

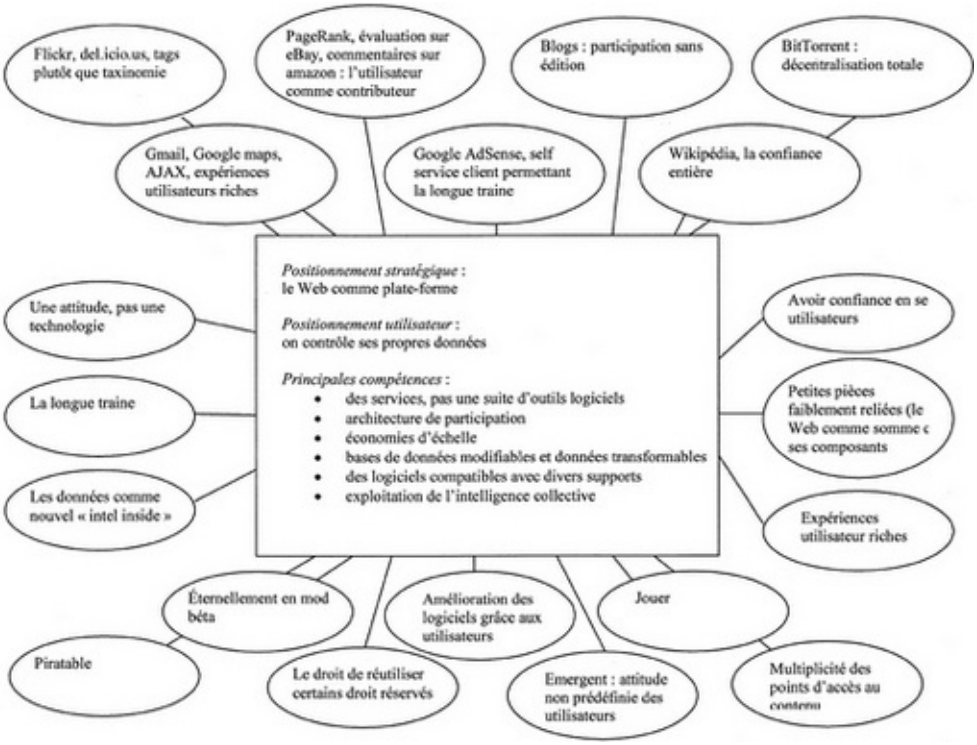


Figure 2. – **Tableau comparatif des sites et des services Web 1.0 et Web 2.0**

Web 1.0	**Web 2.0**
Doubleclick	Google AdSense
Ofoto	Flickr
Akamai	BitTorrent
Mp3.com	Napster
Britannica Online	Wikipédia
Sites perso	Blogs
Evite.com	Upcoming.yahoo.com
Pages vues	Coût au clic
Aspirateurs web	Services web
Systèmes de gestion de contenus	Wikis*
Arborescence (taxonomie*)	Tags* (folksonomie*)
Rigidité du contenu	Syndication de contenu

Figure 3. – **Du Web 1.0 au Web 2.0**

	Web 1.0	**Web 2.0**
Leaders du Web	Entreprises, marchands	Internautes
Profil de l'internaute	Passif	Actif
Interactivité	Sélection et lecture d'information	Sélection, lecture et publication d'information
Unité de recherche	Mot-clé	Tags

d'après ZDNet

Les entreprises et les sites commerciaux du Web 1.0 ont dû abandonner leur position dominante aux internautes dont l'attitude passive est devenue participative sur Internet. Certains sites comme Amazon ont su mettre à profit cette évolution. En 2005, le libraire en ligne a lancé un premier service, Amazon Connect, qui consiste à permettre aux auteurs dont il vend les ouvrages, à créer et maintenir un blog. Ce concept novateur permet pour la première fois aux écrivains de s'adresser directement aux lecteurs à travers des billets publiés sur amazon.com. Les internautes sont à leur tour invités à commenter les livres qu'ils ont lus en postant des commentaires sur shelfari.com qu'Amazon achète en 2007. Il s'agit d'un réseau social dédié aux amateurs de lecture qui stockent sur les étagères d'une librairie virtuelle, les livres qu'ils ont lus ou qu'ils vont lire. Ils attribuent ensuite une note d'appréciation et rédigent une critique qui sera accessible à tous leurs contacts.

À la manière de Facebook, les internautes peuvent créer une liste d'amis et consulter les ouvrages qu'ils ont lus pour partager leurs avis et aussi créer des groupes autour d'un thème ou d'un titre. Amazon a saisi la mouvance du Web 2.0 pour concevoir la plus grande communauté du livre sur Internet. Par ailleurs, Amazon autorise aussi les développeurs à interroger son catalogue depuis d'autres interfaces que celle du son site. Liveplasma.com , anciennement musicplasma.com , propose par exemple ainsi un accès au catalogue d'Amazon *via* une interface riche adossée à un nouveau mode de navigation graphique par affinités stylistiques ou par époque liée au nom recherché.

De la même manière, les nuages de tags assistent la recherche par mots-clés en regroupant un ensemble de termes associés attribués par les utilisateurs en affectant à chacun une taille et une couleur variant selon leur popularité et leur catégorie (personnes, lieux, organisations, appréciations, etc.).

IV. Del.icio.us ou le « social-bookmarking »*

Accessoire indispensable des internautes et plus particulièrement des veilleurs, les favoris permettent de gérer une classification des pages Web dans autant de répertoires que nécessaire. Outre l'obligation d'une rigoureuse organisation, ils présentent l'inconvénient d'être indissociables de l'ordinateur sur lequel ils ont été créés – un véritable handicap lorsque l'on utilise plus d'un poste (un au bureau, un autre à la maison, plus un portable). Sans compter les oublis de sauvegarde qui réduisent à néant, lors d'un crash, les heures passées à classer amoureusement des liens patiemment collectés. Puis il y a le risque principal qui vient tout simplement du fait que l'on a oublié dans quel dossier était conservé tel lien. Bref, les

limites de ce système sont vite atteintes lorsque l'on en fait un usage fréquent parce qu'en toute logique un lien devrait être stocké dans plusieurs répertoires de manière à le retrouver quelle que soit la question que l'on se pose à l'instant précis de la recherche. Mais voilà qu'un des aspects révolutionnaire du Web 2.0 à travers des sites de « social-bookmarking » comme Del.icio.us ont bouleversé à la fois l'ergonomie des favoris mais aussi la recherche en elle-même.

L'utilisateur bénéficie désormais des favoris de tous les internautes du monde, autrement dit toutes les personnes ayant effectué une recherche sur un même sujet vont pouvoir accéder à leurs informations réciproques sur une même plate-forme. Autre avantage inestimable : ces données sont accessibles depuis n'importe quel ordinateur. Le « social-bookmarking » est d'une telle richesse que certains veilleurs n'utilisent plus que ce système pour trouver de l'information et ont abandonné Google et autres traditionnels moteurs de recherche. Les sites comme Del.icio.us fournissent des informations déjà validées et affinées au plus haut niveau grâce aux mots-clés associés. Chacun est libre d'aller visiter les sites remarqués par des dizaines de milliers d'internautes. Il s'agit d'un véritable travail collectif, totalement irréalisable à un niveau individuel. Mais le plus remarquable, dans ce concept, est la possibilité d'aller consulter les favoris des internautes qui partagent nos centres d'intérêt. Del.icio.us indique à chaque tag le nombre d'utilisateurs l'ayant enregistré, ce qui peut être assimilé à une forme d'indice de popularité. Del.icio.us nous prévient même des ajouts de tags effectués par des internautes de notre choix sur un sujet spécifique : si, par exemple, « Marc-Antoine » a un don pour trouver des sites aux contenus originaux sur un thème précis, il suffit de l'enregistrer sur Del.icio.us en précisant son nom et le sujet d'intérêt pour recevoir automatiquement toute nouvelle publication de sa part. Le marquage des sites se fait grâce à un bouton qui ouvre une fenêtre dans laquelle sont inscrits l'url de la page sélectionnée, son titre (modifiable), un champ pour les mots-clés (il est possible d'y inscrire tous les mots auxquels on pense), un autre champ pour des commentaires éventuels (jusqu'à mille caractères) et des tags sont suggérés au cas où l'on serait à court d'idées. Difficile, dans cette perspective, d'oublier où se trouve le précieux lien puisqu'on peut le retrouver dans plusieurs répertoires. Del.icio.us est un système ouvert qui permet d'étoffer son réseau de contacts d'une manière efficace et intelligente en créant des groupes d'utilisateurs qui mettent en commun leurs signets et échangent leurs idées sur les forums.

Attention toutefois à ces systèmes ouverts qui résistent difficilement à l'augmentation du nombre de leurs utilisateurs :

- trop de sites tuent l'effort de qualification : plus il y a de sites, plus il est difficile de se repérer, il faut revenir au moteur de recherche ;

- trop de « users » tuent la notion de folksnomie : à 1 000 personnes, il peut y avoir une certaine homogénéité dans la compréhension et l'attribution du tag, à plusieurs millions d'utilisateurs, de culture et de langue différentes, le cercle n'est plus vertueux, cela devient de la cacophonie ;

- le bruit (commercial, spam, etc.) est un écueil insurmontable : sans « autorité de régulation » ou modérateur, ces systèmes souffrent de leur succès avec l'intrusion de populations plus ou moins bien intentionnées. Et modérer à l'échelle du monde est impossible.

Pionnier du Web 2.0 il y a quatre ans, Del.icio.us a vu sa notoriété et son trafic décroître progressivement au profit des moteurs de recherche et des réseaux sociaux moins anonymes de type Facebook ou autre où je fais confiance à mes amis pour me suggérer des sites, ce qui limite le bruit et les manipulations.

1. Twitter

Twitter est un service de microbloguing permettant aux utilisateurs de poster un message de 140 caractères maximum (soit une ou deux phrases). Selon le principe de fonctionnement du blog, le twitter crée un profil et publie sur la page qui lui est attribuée autant de messages qu'il le souhaite, qu'ils soient personnels ou non, la différence – outre la limitation de la taille du texte – est que les lecteurs ne peuvent pas laisser de commentaire. Ces lecteurs sont en fait des followers (des suiveurs) qui ajoutent l'éditeur du message à leur liste, ce qui permet de recevoir toutes les mises à jour sur leur page personnelle. Au demeurant très simple d'utilisation, l'utilisation de Twitter se révèle assez complexe pour ceux qui veulent en tirer profit, car, pour que l'outil révèle toute sa puissance, il faut déjà bénéficier d'un nombre substantiel de contacts prêts à vous suivre. Par la suite, l'intérêt du contenu publié permettra de se faire connaître (reconnaître), encore faut-il être capable de générer régulièrement des textes succincts dignes d'être lus. En réalité, les power Twitter utilisent cette plate-forme (dont le style télégraphique est idéal) pour lancer des alertes dans lesquelles ils placent un lien, redirigeant ceux qui les lisent vers une page qui, elle, contient l'information. Si le message est intéressant, il sera alors retweeté par les internautes à l'attention de leurs propres followers, ce qui fait de l'outil un système de propagation de signal d'une ampleur jusqu'alors jamais obtenue. L'auditoire étant illimité, les propos peuvent alors avoir un impact phénoménal à un niveau international, au grand dam de la classe dirigeante de pays oppresseurs qui n'arrive pas à contrôler le flux d'information qui transite, minute après minute, par tous les moyens possibles – ordinateurs, téléphones portables, etc. Lors des élections présidentielles en Iran, les mollahs ont enragé de ne pouvoir stopper les milliers de messages et les vidéos postés sur Twitter et sur YouTube par la population, démontrant ainsi que la censure est désormais impossible lorsque les citoyens sont équipés d'outils de communication individuels [12].

Twitter pourrait aussi devenir un outil de contrôle des puissants, des lobbies et autres organisations (voir les événements de l'été 2009).

http://blog.lefigaro.fr/technotes/2009/08/twitter-et-facebook-attaques-la-piste-russo-georgienne.html

C'est une réalité pour tous les services émergents qui deviennent des cibles dès qu'ils sont populaires. Quand une organisation pourra, à coup d'investissements avec des hackers, mettre à terre Twitter, d'autres lieux, sites et modes de communication émergeront.

2. Le Web, c'est aussi l'ère du mouvement permanent, d'une versatilité très rapide des usages

La plate-forme est riche d'informations pour ceux qui auront su se créer les bons réseaux. Toutefois, ces messages doivent mener à des sources vérifiables pour être crédibles ; dans ce cas, l'instantanéité d'une alerte est saisissante. Qu'il s'agisse d'attentats, de conflits, de révoltes ou du décès de célébrités, il ne s'écoule que quelques minutes entre l'événement et la diffusion mondiale de l'information concernant cet événement, au travers des centaines de milliers de « retweets » qui relayeront le premier message publié.

Twitter est, de fait, un complément indispensable pour les sites agrégateurs de news comme Google qui peine à publier l'information en temps réel. Les messages envoyés par les membres de Twitter contiennent parfois des scoops comme la diffusion de la première photo sur l'amerrissage d'un avion de ligne sur l'Hudson River.

Commercialement, Twitter peut être un formidable accélérateur comme l'a si bien compris www.nakedpizza.com qui avoue réaliser plus de 30 % de son chiffre d'affaires grâce à ses followers sur Twitter [13], ou Dell, qui a annoncé avoir vendu pour 3 millions de dollars d'ordinateurs et d'accessoires *via* Twitter depuis son arrivée sur le service de microbloguing, en juin 2007. Les ventes Twitter de la boutique en ligne Dell Outlet, qui solde des ordinateurs retournés par des clients ou rénovés, dépassent ainsi les 2 millions de dollars, mais les microblogueurs ont aussi acheté pour un petit million de dollars de nouveaux ordinateurs.

Dell Outlet revendique être l'un des 50 utilisateurs de Twitter les plus populaires, avec 620 000 suiveurs. Ses commerciaux utilisent le service pour répondre à leurs questions et pour proposer des promotions qui leur sont réservées. Une façon pour le constructeur de gérer rapidement son stock : dès que des articles lui sont retournés, ils sont immédiatement reproposés sur Twitter [14].

Mais le site de microbloguing est aussi victime d'abus, qu'ils soient commis par malveillance, malice ou par jeu, comme l'usurpation d'identité avec, par exemple, la présence de quatre profils de Ségolène Royal. Le premier est surprenant de réalisme, quelques mots peu crédibles mettent la puce à l'oreille mais le nombre de followers atteint presque les 900 personnes, dont la majorité doit être convaincue de lire les messages de la femme politique. Là encore se pose la question de la nécessité d'être présent ou non pour un personnage public. Faut-il éviter de s'inscrire – au risque d'être représenté par d'autres, dont les commentaires pourraient porter atteinte à notre image –, ou, au contraire, être présent, ce qui demande à l'auteur d'y consacrer un certain temps régulièrement, car le follower n'est pas fidèle et, si le contenu se fait rare et/ou n'est pas à la hauteur de l'attente, il se retire et ne revient pas.

3. Plus de bruit que de mots

Plus le nombre d'utilisateurs augmente plus la qualité des messages baisse : on note à présent un accroissement de la pollution par les spams, les fausses rumeurs et les messages autopromotionnels.

Selon une étude réalisée, en août 2009, par le cabinet américain Pear Analytics, plus de 40 % des messages postés par les internautes sur Twitter seraient totalement futiles [15].

2 000 tweets choisis au hasard ont été regroupés en six catégories : actualités, spam, autopromotion, bavardages futiles, conversation et information à faire passer ; il ressort que plus de 40 % de ces minimessages (140 caractères maximum) relèvent du bavardage inutile, 5,85 % de l'autopromotion et 3,75 % du spam. Les conversations entre plusieurs utilisateurs de Twitter ne représentent, quant à elles, que 37 % des tweets. D'où la mise à disposition de nombreuses api* pour filtrer au mieux les flux de données.

V. L'opinion contre l'expertise

Cette forme de pollution est due au succès que remporte le système collaboratif grâce auquel tout le monde peut s'exprimer, ce qui oblige le lecteur à s'interroger sur la qualité des écrits publiés et l'amateurisme de leurs auteurs.

Cette innovation révolutionnaire qu'est le Web 2.0 peut laisser circonspect sur la fiabilité, la traçabilité et la pérennité des informations qu'il véhicule.

Le philosophe allemand et spécialiste des médias, Norbert Bolz, le décrit très justement lors d'un entretien avec le magazine *Der Spiegel* en 2006 : « Ce média cherche encore ses applications. C'est tout à fait normal. On commence par inventer des techniques, puis on réfléchit à ce qu'on peut en faire… On ne peut pas parler de manque de pertinence quand on songe à de nouvelles communautés comme Wikipédia, l'encyclopédie en ligne. On a là tout un savoir de profanes qui entre en concurrence avec le savoir des experts. Pour moi, le mot-clé n'est donc pas démocratisation, mais *doxa*. Les Grecs ont indiqué la voie dans l'Antiquité. Ils ont dit : avant, il y avait la *doxa*, c'est-à-dire l'opinion. À partir de maintenant, nous ne nous intéresserons qu'à la vraie connaissance, à un savoir fondé scientifiquement, l'*épistèmê*. Aujourd'hui, deux mille cinq cents ans plus tard, la *doxa* revient : sur Internet, c'est l'opinion de toutes sortes de personnes qui prévaut, dont très peu sont des experts. Or, en se regroupant, ces opinions offrent des résultats manifestement plus intéressants que ceux des scientifiques hautement spécialisés. C'est cela qui est fascinant avec Wikipédia. Une opinion diffuse et éparpillée rivalise avec le travail universitaire par un étonnant processus d'auto-organisation. […] Le phénomène dissimule également des évolutions économiques très importantes. Une entreprise comme Wikipédia menace l'existence de temples de la connaissance publique comme l'*Encyclopaedia Britannica*. »

Voilà qui est dit : trouver l'information n'est pas suffisant, il faut pouvoir la valider. Si le Web 2.0 est un excellent moyen de tester l'impact d'un nouveau produit ou service sur le marché, il convient de rester prudent, le feed-back des utilisateurs n'étant représenté que par un échantillon de la population qui a souvent plus vite fait de critiquer que d'encenser la toute dernière création. De fait, les informations peuvent souvent être manipulées par une concurrence déloyale, d'où la nécessité d'un comparatif : gain potentiel et risque.

VI. Le renseignement étatique par le Web 2.0

Fin octobre 2008, l'armée américaine publiait un rapport sur l'utilisation des technologies de communication par des organisations terroristes. Cette étude, fondée sur des sources ouvertes accessibles sur Internet, fait état des nombreux moyens de communication dont disposent aujourd'hui les « hacktivistes », terroristes, militants politiques, anarchistes et autres, à travers des outils comme Skype, Twitter, le gps, le mashup*, la géolocalisation, etc. Ce que l'on appelle l'Open Source intelligence (osint) est en voie de révolutionner la pratique du renseignement militaire. Parce que le Web 2.0 est accessible aux terroristes et aux organisations criminelles, il devient nécessaire à la communauté du renseignement de comprendre ses fonctionnements et d'en exploiter à son compte la technologie.

L'odni (Office of the Director of National Intelligence) a lancé Intellipedia en 2006 qui utilise le même logiciel que Wikipédia (Media Wiki). Un système collaboratif qui consiste à relier les analystes, les experts, les groupes de travail et les collecteurs de données, puis à permettre aux utilisateurs d'afficher, modifier ou améliorer les articles.

Deux ans après son lancement, Intellipedia compte plus de 330 000 pages, 42 204 utilisateurs inscrits (toutes catégories d'âge confondues) et quelque 135 000 lecteurs. Cette croissance est tout simplement spectaculaire. Intellipedia a atteint le million de contributions en deux mois de moins que Wikipédia.

Mais Intellipedia est plus qu'un dépôt d'informations ; c'est désormais un outil qui permet aux agents de renseignements du monde entier d'analyser les crises à mesure qu'elles éclatent. À une époque où cnn diffuse des reportages à rebondissements et où les citoyens regardent les dernières nouvelles défiler au bas de leur écran d'ordinateur, Intellipedia se déclare comme un outil puissant qui permet d'échanger des

informations et d'analyser les incidents planétaires pratiquement en temps réel [16].

La communauté du renseignement est en train de réaliser l'intérêt que représente le renseignement « open source » qui permet de valider des informations grises en économisant énergie et argent par la collecte, le croisement et l'analyse des données récupérées. Mais aussi d'utiliser les réseaux de compétence et d'analyse collaborative afin d'en finir avec le travail individuel et d'élargir les compétences et les ressources de la communauté en tablant sur de meilleures collaborations avec les partenaires et les clients extérieurs au service (publics et privés). Selon le rapport, les frontières et lignes de démarcation géographiques, juridiques et conceptuelles s'estompent, se brouillent et il devient, dès lors, plus difficile de distinguer le renseignement de l'information, le privé du public, les alliés des concurrents [17].

À l'horizon 2015, le renseignement américain aura ainsi adopté une démarche collaborative globale, qui intégrera des compétences variées dans le cadre de synergies dédiées à la réalisation de missions, selon une logique de relation client (Customer relationship) [18].

Notes

[1] Abram Shulsky et Gary Schmitt , *Silent Warfare: Understanding the World of Intelligence*, Washington DC, Brassey's, 3e éd., 2002.
[2] Roger Faligot et Rémi Kauffer , *Histoire mondiale du renseignement* t. I, : *1870-1939*, Paris, Robert Laffont, 1993.
[3] Source : www.secretservice.gov
[4] Pierre Conesa, ceis, www.diploweb.com/forum/renseignement 36111.htm
[5] Source : http://echelononline.free.fr/
[6] Source : www.cncpi.fr
[7] Bernard Besson , *Du renseignement à l'intelligence économique*, Paris, Dunod, 2e éd., 2001.
[8] L'Atelier bnp Paribas.
[9] Chiffres communiqués en janvier 2009 par le cabinet ComScore spécialisé dans la mesure d'audience.
[10] La première occurrence des mots suivis de * renvoie au glossaire en fin d'ouvrage.
[11] Kevin Kelly, « The Wired ».
[12] Source : http://www.lemonde.fr/opinions/article/2009/06/25/twitter-la-crise-iranienne-et-les-mobilisations-citoyennes-par-yves-mamou_1211292_3232.html
[13] Source : http://www.thekbuzz.com/casestudies/Twitter
[14] Source : http://www.informaticien.be
[15] Source : lemonde.fr
[16] Michael Wertheimer, sous-directeur adjoint, technologie et transformation analytique, Bureau du directeur du renseignement national des États-Unis, « Armer le renseignement avec le Web 2.0 ».
[17] Rapport John M. McConnel, *« Vision 2015: A Globally Networked and Integrated Intelligence Enterprise »*, dni, juillet 2008.
[18] Franck Bulinge , *« Renseignement et Intelligence économique »* , blog d'études et de recherches.

Chapitre II

Histoire des moteurs de recherche

Au fur et à mesure de l'inexorable développement d'Internet, les internautes sont confrontés à un problème croissant : comment retrouver l'information recherchée dans la plus grande bibliothèque du monde ? Le seul moyen de retrouver des données dans cette masse d'information serait de référencer l'intégralité des sites Web, de les indexer et de restituer les données selon une architecture organisée pour une meilleure compréhension du lecteur.

À partir de cette réflexion sont nés les premiers moteurs de recherche, dont Archie en 1990 et Wanderer en 1993. Mais le premier annuaire, né de l'esprit de deux étudiants californiens de l'université de Stanford, Jerry Yang et David Filo, fut Yahoo ! qui connut un succès immédiat. Son activité principale était d'indexer manuellement les sites Web.

Puis sont apparus, Lycos et Excite en 1995. AltaVista, qui signifie « vue d'en haut », est née d'un ensemble de grandes idées d'une équipe d'experts obsédés par l'information. Au cours du printemps 1995, des scientifiques du laboratoire de recherche en informatique de Palo Alto, en Californie, ont imaginé une méthode de stockage dans un index de recherche rapide de n'importe quel mot issu de n'importe quelle page html d'Internet. Cela conduisit au développement par AltaVista de la première base de données de recherche de texte intégral du monde sur le Web et fut aussi le premier moteur de recherche multilingue.

En 1998, Google révolutionne le concept de la recherche en ligne grâce à son moteur dont l'interface dépouillée (brevetée depuis sept. 2009) séduit la majorité des internautes. Selon le site Gawker, le moteur de recherche a obtenu de l'Agence fédérale américaine un brevet pour sa page d'accueil, présentée comme une « interface utilisateur graphique d'un terminal de communication pour écran ». Le moteur de recherche pourrait utiliser ce document pour lutter contre certains concurrents qui seraient tentés de présenter une interface similaire. À la fois pertinent et exhaustif, Google se singularise par son système de référencement fondé sur la popularité des sites auprès des internautes. Rapidement, le groupe américain passe en tête de tous les moteurs de recherche.

En réalité, l'histoire a commencé en 1995 lors d'une visite d'étudiants à l'université de Stanford pendant laquelle Larry Page et Sergey Brin se sont rencontrés. Partageant une vision commune du moteur de recherche idéal, ils créent, en 1996, BlackRub, un premier moteur hébergé sur une machine constituée d'un boîtier à base de Lego® et de 10 disques durs de 4 gigaoctets chacun [1]. Trois ans plus tard, Google est lancé et doit une partie de son immense succès au fameux PageRank – un système de classement unique basé sur l'indice de popularité d'une page. Il détermine l'ordre et la pertinence des liens dans les résultats de recherche qu'il fournit. Il faut bien reconnaître qu'avant le PageRank, les résultats des moteurs de recherche étaient relativement aléatoires et une même requête sur un même moteur pouvait présenter des résultats différents. Le système de Google s'est donc révélé le plus fiable. PageRank permet de mesurer objectivement l'importance des pages Web. Ce classement est effectué grâce à la résolution d'une équation de plus de 500 millions de variables et de plus de 2 milliards de

termes. Au lieu de compter les liens directs, PageRank interprète chaque lien de la Page A vers la Page B comme un vote par la Page A pour la Page B. PageRank évalue ensuite l'importance des pages en fonction du nombre de votes qu'elles reçoivent [2].

PageRank tient également compte de l'importance de chaque page qui « vote » et attribue une valeur supérieure aux votes émanant de pages considérées comme importantes. Les pages importantes bénéficient d'un meilleur classement PageRank et apparaissent en haut des résultats de recherche. La technologie de Google utilise l'intelligence collective du Web pour déterminer l'importance d'une page. Les résultats ne font l'objet d'aucune intervention humaine ni manipulation, ce qui explique pourquoi les internautes font confiance à Google et considèrent ce moteur de recherche comme une source d'informations objective et indépendante.

Google a aussi été le premier à s'ouvrir aux développeurs du Web en leur permettant de comprendre comment et où ils se positionnaient sur la Toile pour qu'ils puissent travailler leurs sites de manière à être indexés par le moteur de recherche. Le moteur a judicieusement choisi d'autoriser les webmasters à développer des sites qui seraient plus faciles pour lui à indexer que l'inverse, ce qui a considérablement contribué à sa popularité. Google consacrerait 50 % de ses ressources humaines en recherche et développement (R&D) pour fournir aux internautes des services totalement novateurs comme Google Earth, Google Maps, Street View ou Picasa.

À la même époque, en juillet 1998, France Télécom, par son département Internet Wanadoo, lance le moteur Voila, positionné comme un concurrent de Yahoo ! plus que de Google. Cependant, étant installé par défaut sur les boîtiers de connexion à Internet de Wanadoo, il est largement utilisé en France. Voila indexe aujourd'hui environ 60 millions de pages en français.

Fast Technology s'inspire du succès de Google pour créer en 1999 le moteur AllTheWeb et réserve son moteur Fast aux entreprises dont la masse d'informations qu'elles détiennent nécessite une organisation structurée.

Dans le grand public, les usages sont polarisés sur deux ou trois moteurs américains (Google, Yahoo !, Msn) et sur Voila, mais, parmi les propositions émergentes, Exalead est un moteur français souvent utilisé par les veilleurs, les acteurs de l'intelligence économique, les chercheurs et les étudiants pour aller chercher de l'information différente, mais aussi de plus en plus par le grand public, parfois désemparé par l'affichage brut que présentent ses concurrents.

Exalead, fondé en 2000 par Patrice Bertin et François Bourdoncle [3], fait son apparition en 2006, après avoir lancé sa version bêta en 2001. Ce moteur généraliste, conçu en France, utilise une technologie basée uniquement sur des algorithmes statistiques qui peuvent traiter des corpus très importants et permettent de proposer des fonctions intelligentes telles que la recherche sémantique multilingue ou le choix du format d'affichage des résultats avec ou sans vignette de prévisualisation. Mais là où Exalead se distingue, c'est dans l'assistance à la recherche grâce à son outil de recherche affinée qui s'affiche à droite de la page de résultat. Cet outil offre différents filtres permettant d'optimiser les recherches : termes associés, langues, zones géographiques, types de sites (blog, forum, commercial ou non commercial), multimédia, etc.

En France, en mars 2009, plus de 89 % (89,57 %) des visites moteurs générées le sont *via* Google, loin derrière se trouve LiveSearch avec 2,84 % talonné par Yahoo ! avec 2,51 % de visites. Aol est en 4^{e} position avec 1,76 % et Orange arrive le 5^{e} avec 1,58 % des consultations [4]. La lecture de ces chiffres

doit tenir compte des « Search bars » qui sont intégrées par défaut soit aux navigateurs (LiveSearch pour Internet Explorer), soit qui s'exécutent automatiquement lors du téléchargement d'une application tierce qui propose l'installation d'une toolbar (Google toolbar ou Yahoo ! toolbar), ou encore qui sont installées par défaut lors de la mise en œuvre du boîtier de connexion à Internet (Orange avec la LiveBox). La majorité des utilisateurs vont rester bloqués sur un seul système. Il serait intéressant de connaître le pourcentage de ces visites « par défaut » qui doivent sensiblement fausser les résultats.

L'expansion extraordinaire d'Internet a parallèlement développé le comportement des internautes en matière de recherche d'informations. Les utilisateurs ont désormais un niveau d'exigence élevé sur l'architecture des données restituées par les moteurs. Mais l'observation comportementale démontre des failles dans le process des outils de recherche. En effet, selon une étude coréalisée par Jupiter Research et iProspect [5] en 2006, 62 % des internautes cliquent sur un résultat affiché sur la première page du moteur consulté, 19 % de plus se rendent sur la 2e page et encore 9 % vont jusqu'à la 3e. Soit 9 utilisateurs sur 10 cliqueront sur un lien contenu dans les 3 premières pages des résultats affichés.

Le niveau de recherche est donc limité d'autant, ce qui, dans le cas de Google, tend à leurrer l'internaute, compte tenu du fait que le classement se fait en fonction de la popularité d'un site, si celui-ci figure dans le top trois des pages, il sera largement plus consulté que les autres puisque les utilisateurs se contentent, dans la majorité des cas, des trois premières pages affichées. De ce fait, le site s'assure une position renforcée sur les moteurs de recherche. Cette méthode implique également que le contenu du site, sa valeur et celle de l'information qu'elle contient ne sont pas pris en compte. Un moyen efficace de diffuser de l'information erronée sachant qu'elle sera reprise par 90 % des internautes [6].

Aujourd'hui, nous voyons apparaître de faux moteurs de recherche, conçus par les organisations criminelles dont les résultats affichés dirigent les internautes vers des pages infectées par des virus ou des chevaux de Troie, notamment sous la forme de logiciels players, permettant soi-disant de visualiser des vidéos ou autres fonctionnalités. Un de ces outils, détecté par l'éditeur d'antivirus PandaSecurity, aurait déjà été utilisé par près de 200 000 personnes [7].

I. Intégration du Web 2.0 dans les moteurs de recherche

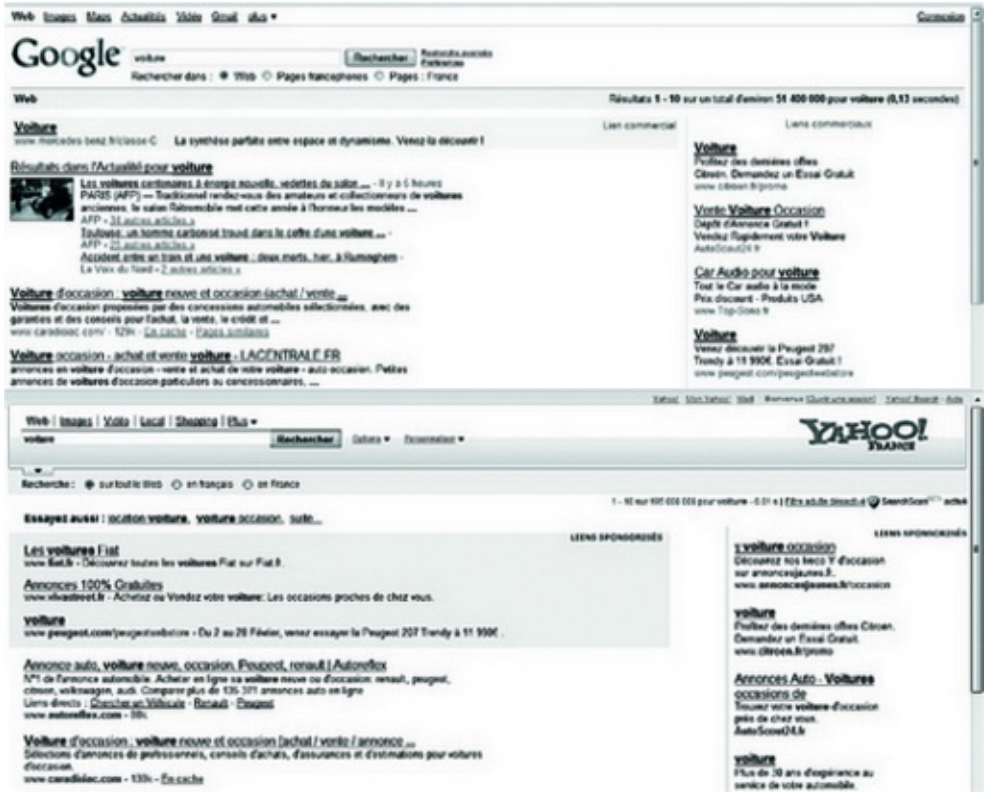
Conscients de l'émergence grandissante de ce nouveau concept, les moteurs de recherche ont intégré la technologie de Web 2.0 en offrant aux internautes de nouveaux services développés en interne ou en rachetant des sites populaires. Ce fut le cas pour Google qui offre des services comme Google Maps et a racheté YouTube, le site de partage de vidéos. Yahoo ! pour sa part propose Yahoo Maps, Yahoo Answers ou MonWeb 2.0 pour créer sa base de sites à conserver et classer à l'aide de tags et a racheté Flickr et Del.icio.us en 2005.

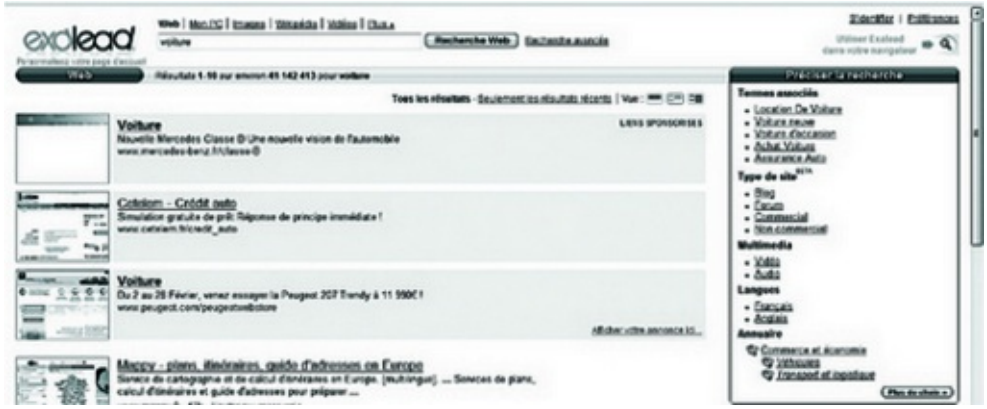
Restminer, développé par Exalead, est un autre exemple de cette innovante forme de recherche. Il offre une vision dynamique mêlant du contenu généré par les utilisateurs sur les blogs, affiche un « buzzranking » des internautes, une géolocalisation grâce à l'api Google Maps et des informations issues de sources diverses du Web et de Wikipédia. L'internaute est alors entièrement assisté pour parvenir à formuler sa requête et obtenir une remontée d'informations cohérente.

II. Étude comparative

1. 1re recherche

Pour mieux comprendre ce fonctionnement, une simple recherche sur le mot « voiture » a été réalisée sur les moteurs de recherche suivants : Google, Yahoo ! et Exalead.





Ce test permet de constater que les résultats sur Google comme sur Yahoo ! sont bruts avec toutefois un nombre de pages. La colonne de droite est consacrée à des liens commerciaux alors qu'Exalead propose dans son menu Exaleader des filtres destinés à affiner la recherche. L'internaute peut alors associer sa recherche à des termes auxquels il n'aurait pas initialement pensé. Toujours dans le cadre de la recherche sur le mot « voiture », Exalead suggère la location, la vente ou l'achat de voitures, mais aussi des recherches ciblées sur des forums, des blogs ou des sites commerciaux, en français ou en anglais. Un bouton « plus de choix » développe l'Exaleader pour affiner les critères et précise les pourcentages de résultats pour les langues, types de documents et situations géographiques.

2. 2e recherche

Une requête plus précise d'un internaute sera également interprétée différemment par les moteurs de recherche testés. Ainsi la demande « resto paris magret de canard » présentera les résultats suivants :

Si les moteurs Google et Yahoo ! remontent des informations pertinentes, aucun moyen d'étendre la recherche n'est proposé. De son côté, Exalead offre de nouveau une recherche affinée par des critères suggérés dans son menu Exaleader. Ces critères peuvent être sélectionnés, supprimés ou exclus selon les souhaits de l'internaute. L'Exaleader est situé à droite de l'écran et propose des critères spécifiques à la restauration puisque la recherche de l'utilisateur porte sur des restaurants. Il peut donc définir le lieu géographique, la catégorie de restaurant, indiquer une fourchette de prix, un type de cuisine, d'établissement (gastronomique, brasserie, hôtel, etc.), l'environnement, les horaires d'ouverture et les distinctions dont bénéficierait le restaurant. À gauche, une carte Google Maps permet de localiser le restaurant sélectionné en survolant son nom avec la souris.

C'est dans ce cadre, précisément, que se distinguent la qualité et la pertinence d'un moteur de recherche. En partant de l'analyse de Jupiter Research et iProspect selon laquelle 90 % des internautes se contentent

des résultats affichés dans les trois premières pages, il devient presque aisé d'en conclure que les utilisateurs ne savent pas chercher l'information.

La recherche d'informations sur des sujets aussi sensibles et controversés que le nucléaire nécessite une réflexion plus poussée. Cependant, la difficulté reste la même : comment formuler correctement sa requête ?, voire, que cherche-t-on exactement ? Là encore, le système assiste l'internaute et lui suggère de préciser sa recherche selon que l'on désire se renseigner sur des centrales, l'énergie, les déchets, les réacteurs, les accidents et bien d'autres thèmes encore. Si le renseignement concerne l'opinion des citoyens ou des associations écologiques, il faudra sélectionner « blog » ou « forum » dans le type de site. Il est possible d'affiner plus encore la recherche et de choisir le type de fichier, pdf, Word, Text, Excel, etc. Ou encore de retrouver toutes les vidéos disponibles sur le Net. Si l'on ajoute la sélection géographique, il devient possible de trouver très précisément l'information désirée, puisque le nombre de pages remontées diminue (en toute logique) proportionnellement au nombre de critères ajoutés à la requête initiale.

Ce qui peut désorienter un instant l'internaute, habitué depuis quelques années à juger de la pertinence d'un moteur de recherche en fonction du nombre de pages remontées, est rapidement compensé par l'attrait de voir enfin un outil capable de comprendre ses besoins réels. Avoir une requête dont le nombre de réponses atteint quelques millions de pages tend à rassurer les internautes, même si 10 % à peine d'entre eux dépasseront la 3^{e} page et moins de 1 % la 10^{e}. Quel intérêt, dès lors, d'avoir un moteur qui en propose 10 millions ? Probablement un besoin psychologique similaire à celui qui invite à remplir son réfrigérateur pour ne manquer de rien, au cas où une petite faim se ferait sentir, la quantité primant sur la qualité. L'internaute se sent perdu avec trois pages ciblant exactement sa recherche, et quasi rassasié avec 3 000, qu'il ne lira jamais. Mais le nombre donne un sentiment d'avoir accompli son travail, d'avoir fait le tour de la question. Tout cela n'est donc, au final, qu'un leurre. En conséquence, l'ultime moteur de recherche est bien celui qui accompagne son utilisateur en lui procurant des moyens concrets sur les termes qu'il pourrait associer à sa requête pour aboutir au résultat souhaité.

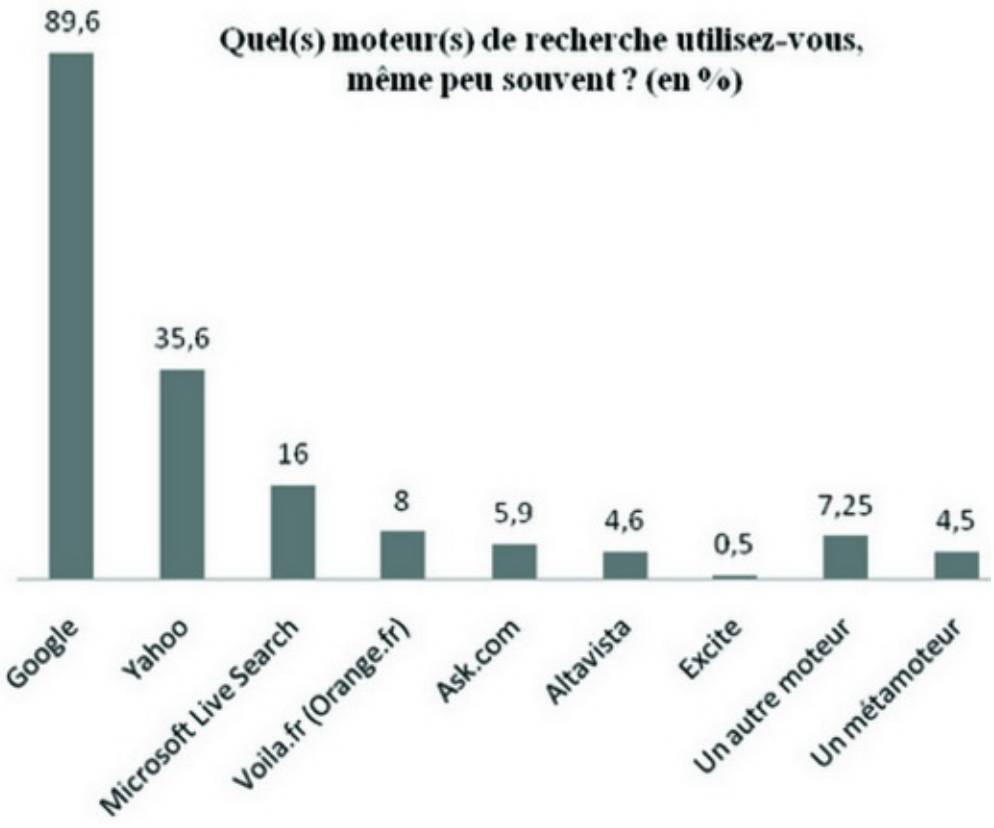


Figure 10. – **Indice de popularité des moteurs de recherche en France**

Source : enquête du *Journal du Net*, mai 2009.

Notes

[1] Source : http://www.rankspirit.com/histoire-google.php
[2] Source : http://www.google.fr/intl/fr/corporate/tech.html
[3] Ingénieur de l'École des mines de Paris, François Bourdoncle a participé à la création du moteur de recherche AltaVista, vedette des moteurs de recherche dans la seconde moitié des années 1990 (aujourd'hui technologie propriété de Yahoo !). Il y avait développé la fonction Refine qui permettait d'affiner une recherche en proposant plusieurs mots-clés proches par analyse statistique du contenu des pages trouvées par la requête.
[4] Source : at Internet Institute.
[5] Source : www.wmaker.net : « chercher c'est trouver ».
[6] C'est ce qui deviendra plus tard les Joe Jobs et autres Google Bombing (voir chapitre IV, III : Les dérives criminelles).
[7] Source : abondance.com

Chapitre III

La veille

La veille est le processus par lequel l'entreprise recherche des informations concernant l'évolution de son environnement socio-économique pour créer des opportunités de développement et réduire ses risques face à la concurrence et à l'évolution du marché.

Jusqu'aux environs de l'année 2005, la veille s'effectuait (en France) surtout sur les cinq thèmes suivants, puis l'intégration d'Internet dans les foyers et les entreprises, liée à l'évolution des plates-formes technologiques, a largement modifié à la fois les méthodes et les thèmes.

I. Les différents types de veille

1. La veille juridique

Recherche, analyse et mise à jour de toutes les informations d'ordre juridique et plus particulièrement la jurisprudence, les lois et les décrets, les propositions de loi. La mise en place de cette veille est notamment liée au souhait de pénétration d'une entreprise sur un marché étranger, mais elle peut aussi concerner le marché national.

2. La veille concurrentielle

Cette veille concerne toute information relative à la concurrence. Analyse des produits ou services, leurs ressources humaines et financières, leur stratégie, leur communication, leurs tarifs. Objectif : classement général des concurrents par ordre de performance globale pour identifier les plus menaçants. Amélioration de la communication et de la stratégie et éventuellement de la politique tarifaire. Évolution de la gamme des produits et/ou services en fonction des tendances du marché concurrentiel. Identification des clients et mise en place d'une stratégie en vue de récupérer des parts de marché. Recrutement des meilleurs éléments.

3. La veille technologique

Il s'agit de l'observation et de l'analyse des informations ayant trait aux acquis scientifiques, techniques, à la recherche et au développement, aux produits, aux procédés de fabrication, aux matériaux et aux dépôts de brevets pour en déduire les opportunités de développement ou les menaces. Cette veille permet en outre de réduire considérablement les frais de recherche et développement en interne.

4. La veille marketing

Suivi et analyse de nouveaux marchés. Recherche de signaux plus ou moins forts émis par la clientèle de

façon à détecter l'évolution des attitudes, besoins ou attentes des consommateurs et ce dans le but de leur proposer de nouveaux produits ou services. Au-delà de la collecte d'informations, il faut se rendre dans les endroits branchés et sentir les nouvelles tendances pour innover et mettre à jour continuellement sa communication. Depuis le Web 2.0, la veille marketing peut se faire pratiquement exclusivement sur le Net.

5. La veille stratégique

Elle reprend les grandes lignes des veilles ci-dessus. Ce concentré d'informations offre une vue globale de l'environnement professionnel. C'est d'ailleurs en fonction des questions que l'on se pose lorsque l'on souhaite créer une cellule de veille, que l'on pourra déterminer les équipes de travail à mettre en place pour obtenir une information stratégique et la rediffuser aux acteurs décisionnaires. Si certains d'entre eux désirent une information ciblée, elle sera l'élément mis en valeur au sein de l'information stratégique globale, mais ne pourra en aucun cas être isolée. Les risques et les opportunités ne peuvent se mesurer qu'à l'aide d'une information complète.

La veille sur Internet implique une connaissance approfondie des fonctionnalités des moteurs de recherche et des agents de veille disponibles sur la Toile. Ils possèdent chacun des particularités et si 95 % des résultats sont identiques, les 5 % restants valent largement le temps consacré au tri effectué. Cela dit, il est tout aussi nécessaire de faire une mise à jour manuelle des informations. En premier lieu, il convient d'identifier clairement en quoi consiste la recherche, ensuite il faut :

- définir le secteur d'activité ;
- identifier les acteurs concernés ;
- lister des fournisseurs, concurrents, salons professionnels, etc. ;
- lister les mots ou les expressions clés.

Cette étude définira le point de départ de la recherche qui s'affinera jour après jour en fonction des résultats obtenus.

Tous les documents mis en ligne sont précieux, il faut toujours les télécharger pour récupérer des données cachées. Les fichiers propriétaires comme Word (Microsoft) ou pdf (Adobe) sont riches d'informations dont les auteurs paient au prix fort leur découverte. Exalead propose d'ailleurs une recherche par type de document (pdf, Word, PowerPoint, etc.).

Accéder aux propriétés d'un fichier Word permet d'obtenir des détails qui peuvent se révéler surprenants :

- l'auteur du document d'origine (le document peut avoir été subtilisé à un concurrent) ;
- la date réelle de sa création (idéale pour s'assurer que la proposition commerciale « sur mesure » est en fait un standard de l'entreprise) ;
- le dossier dans lequel il se situe sur l'ordinateur (C :Documents and SettingsAlbert DupontClients difficilesLacible S.AProposition.doc) ;

- l'annulation des modifications permet de revenir aux versions antérieures.

Quant aux fichiers au format pdf, le logiciel Acrobat version Pro d'Adobe permet de bénéficier d'une palette d'outils aux fonctionnalités indispensables lorsqu'un document doit être lu et commenté par plusieurs services. Chacun peut ajouter des annotations et suggérer corrections de texte ou de mise en page. Le fichier finalisé est envoyé au destinataire qui l'ouvrira avec l'utilitaire Acrobat Reader (disponible gratuitement sur Internet). Cependant, rien n'affirme que l'entreprise qui doit réceptionner le document ne soit pas aussi équipée de l'application Acrobat Pro ; dans ce cas, toutes les annotations seront visibles sur l'ordinateur récepteur.

En mai 2006, un document au format pdf, publié sur le site Internet d'un gouvernement, a été téléchargé. Ouvert avec Acrobat Pro, il a laissé apparaître le nom d'une des personnes ayant participé à la rédaction du rapport. Une recherche sur Internet a permis de récupérer la photo de cette personne, ses coordonnées, son parcours professionnel et ses contacts (noms, prénoms, professions) au sein des différents ministères de ce gouvernement. Dans une opération d'approche d'une cible par le Social Engineering, cette méthode est redoutable d'efficacité.

Ces formes de veille, plutôt classiques, sont aujourd'hui reconsidérées grâce aux nouvelles plates-formes du Web 2.0 : les utilisateurs peuvent désormais obtenir un panorama particulièrement précis sur l'ensemble des informations les concernant. Trouver l'information pertinente sur Internet demande une parfaite connaissance des différents outils de recherche et de leur fonctionnement.

Les flux rss* intégrés dans la plupart des sites de news, d'entreprises, les blogs ou les sites de partages offrent la possibilité de recevoir des informations issues de nombreuses sources impossibles à surveiller manuellement.

Les réseaux sociaux comme Facebook, MySpace ou LinkedIn, les sites « Digg like » qui recueillent des actualités recensées et postées par les utilisateurs sont des sources riches en informations. Mais il convient aussi d'étendre ces recherches à des sites de services plus spécifiques : Flickr est un site de partage d'images qui permet de mettre en ligne ses photos qui peuvent être libres d'accès ou réservées exclusivement à ses proches. La recherche se fait par mots-clés et peut être orientée par les tags attribués par les utilisateurs pour les définir. Elle permet alors de visualiser des informations sur une entreprise : ses locaux, ses dirigeants, les salons auxquels elle participe, ses produits, etc. Mais surtout d'accéder à des photos de manifestations contestataires ou de logos détournés. Les sites de vidéo comme YouTube ou DailyMotion ou les sites de partage de favoris comme Del.icio.us permettent de compléter les recherches et d'obtenir un tableau assez exhaustif sur un sujet déterminé.

Pour trouver les sites spécialisés dans ces services, les moteurs comme Google ou Exalead intègrent désormais des recherches sur les supports multimédias. Exalead propose aussi une recherche audio qui donne accès, entre autres formats mp3 ou mp4, aux podcasts téléchargeables sur Internet.

Les entreprises, lorsqu'elles n'ont pas de salarié dédié à cette activité, font alors appel à des professionnels de la veille pour obtenir une analyse fine de l'ensemble des informations qui les intéressent. Il existe beaucoup d'entreprises spécialisées dans la veille (particulièrement dans les officines privées d'intelligence économique), la différence se fait au niveau de la qualité des services proposés et du professionnalisme des analystes. Cybion [1], un des pionniers de la veille technologique en France, créé en 1996, articule ses recherches autour de quatre déclinaisons dont les sources s'incrémentent en fonction de l'émergence de nouvelles sources identifiées :

- la veille à partir de la presse en ligne ;
- la surveillance d'Internet formel qui comprend le suivi de la presse ;
- la surveillance d'Internet informel ;
- la surveillance de l'ensemble des informations disponibles sur le Net.

Au fil du temps, l'entreprise s'est structurée en trois unités hautement spécialisées et un laboratoire :

- une *unité finance* chargée d'identifier les bruits et rumeurs dans le domaine de la banque, des assurances et de la Bourse. Une *unité luxe* dédiée au développement de la veille de réputation et l'anticipation de crise dans le secteur du très haut de gamme où les entreprises sont fortement dépendantes de leur image grand public aussi bien en France qu'à l'international ;
- une *unité développement durable et santé* qui suit les problématiques liées à la gouvernance d'entreprise ainsi qu'à l'environnement, l'écologie et la santé, avec un soutien substantiel dans le domaine de la pharmacie et des cosmétiques ;
- le *laboratoire R&D* teste et compare l'ensemble de l'offre logicielle actuelle dont les résultats sont régulièrement publiés sur le site www.veille.com. Pour Cybion, la veille en tant que telle ne représente plus que 40 % de ses ressources, ces dernières étant désormais focalisées sur celles offrant une haute valeur ajoutée, comme l'analyse stratégique qui permet de mettre en perspective les informations récupérées et traitées, la fiabilité des sources, le recoupement d'information, l'impact pour le client, le swot [2], etc.

Samuel Morillon (directeur général de Cybion), dans une interview du mois de novembre 2008 (France Info), précise que « 85 % du Web est généré par les internautes. Dans cette fange numérique, le grand enjeu est de se remettre dans la conversation des 9 millions de blogueurs français [3] pour anticiper d'éventuelles situations de crise mais aussi pour saisir des opportunités et récupérer des idées. En effet, les frictions opposent des entreprises organisées à des communautés d'internautes diffuses ou à des entreprises entre elles *via* des communautés numériques interposées. Les affrontements, devenus asymétriques et décentralisés, conceptualisent le retour de l'anarchie ».

II. L'e-réputation

La réputation se crée et se propage sur Internet. Et chaque internaute est un média potentiel.

Désormais, marques, produits, entreprises ou collaborateurs peuvent être le sujet de conversations, d'attaques mais aussi de rumeurs positives relayées sur Internet. En effet, tout un chacun peut désormais s'exprimer facilement *via* des outils simples en diffusant sa propre information ou en commentant celle des autres et donc participer à la réputation d'une personne ou d'une organisation.

La gestion de la réputation est pourtant l'un de ces domaines qui est souvent laissé de côté jusqu'au dernier moment. Savoir ce qui se dit sur soi, sur son site ou ses produits est essentiel dans le monde d'Internet où l'on doit maîtriser son identité et sa réputation numériques. « Si vous rendez vos clients mécontents dans le monde réel, ils sont susceptibles d'en parler chacun à six amis. Sur Internet, vos clients mécontents peuvent en parler chacun à 6 000 amis », explique Jeff Bezos, le p-dg de la célèbre

boutique en ligne Amazon :

« L'internaute peut devenir ainsi producteur de contenu : *via* l'écriture ou l'ajout de commentaires, articles dans un blog, un wiki ou un site de presse collaboratif.

L'internaute peut également être organisateur de contenu : en créant les rubriques d'un wiki ou d'un blog, en chargeant ses photos ou vidéos sur une plate-forme multimédia et en l'identifiant *via* des tags (c'est-à-dire en collant une étiquette constituée de mots-clés sur son contenu afin de le définir).

L'internaute sera aussi souvent diffuseur d'information : en écrivant sur son propre blog, en commentant les billets des autres blogs ou les articles de journalistes, en publiant ses photos et vidéos, il diffuse en effet l'information de son choix, au plus grand nombre. Elle peut être anecdotique et inoffensive comme violente, offensive et porter préjudice à d'autres individus ou organismes. Par ailleurs, *via* le flux rss, l'information peut se démultiplier et se diffuser très rapidement, rendant difficile, par exemple, l'identification de la source originale ou le retrait d'un article dérangeant. » [4]

Quasiment impossible à combattre, sauf par certains professionnels de la gestion de crise, la jouissance de la nuisance en ligne est telle que les plus exposés, à savoir ceux dont les produits ou les services touchent en priorité les communautés d'internautes, renoncent à lutter contre les écrits nauséeux qui jaillissent en permanence sur la Toile. Ainsi, le site 97thfloor.com constate que pratiquement toutes les entreprises du *Fortune* 100 sont référencées par centaines sur Google si l'on ajoute « sucks » (pue) après avoir saisi leur nom sur le moteur de recherche.

Par ailleurs, il existe aussi des avantages à consulter les sites sur lesquels les internautes s'expriment librement, le principal d'entre eux étant d'obtenir en temps réel la température du consommateur sur un produit ou un service. Lorsque l'utilisateur est mécontent et le manifeste assez clairement pour porter un préjudice significatif à une entreprise, celle-ci doit répondre dans les règles de l'art et en respectant l'intelligence du client. La société Belkin (fabricant de périphériques informatiques) en a fait l'expérience sur l'un de ses produits dénigré par nombre d'utilisateurs. Elle a donc offert aux internautes à travers le site d'amazon.com d'être rémunérés à hauteur de 0,65 dollar, pour chaque avis positif posté. La référence de l'article en question était fournie avec les liens vers les sites où publier les contributions. Rapidement confondu, Belkin a dû présenter ses excuses et retirer toutes les traces de cette pitoyable tentative de corruption [5].

Pourtant à la limite de la légalité, la manipulation du consommateur sur Internet n'est plus un accident : c'est un secteur d'activité à part entière, nommé l'Astroturfing, qui décrit l'activité d'agents d'influence du Net. Des internautes contribuent à des sites, forums ou blogs en publiant des commentaires qu'ils tentent de présenter comme une participation spontanée.

III. L'e-réputation des États ou la corruption du buzz numérique

Les internautes et les entreprises ne sont pas les seuls à profiter du buzz digital ; ainsi, selon l'ancien correspondant du *Herald Tribune* et du *New York Times*, Thomas Crampton [6], le gouvernement chinois manipulerait l'opinion sur Internet à travers ce qu'il nomme les « 50 cts agents » ou les agents d'influence à 50 centimes de Renminbis [7] (0,05 euro) pour chaque commentaire favorable au gouvernement posté sur Internet.

Le journaliste a interviewé une consœur à Hong Kong qui affirme que ces free-lances engagés par les autorités chinoises arrondissent leurs fins de mois grâce à ces interventions en ligne. Les commentaires propagandistes auraient commencé courant 2007 et auraient atteint un point culminant lors des préparatifs des Jeux olympiques de 2008, lorsque la Chine était à l'époque particulièrement ciblée par la communauté internationale, notamment pour sa politique des droits de l'homme. Les commentaires seraient, à l'instar du modèle chinois en matière de communication, destinés à influencer en douceur sans entrer en conflit avec les auteurs des propos contestataires. Extraits de l'interview de Thomas Crampton :

« Qui sont les "50 cents people" en Chine ?

Ce sont des free-lances qui tiennent des blogs. Ils travaillent pour le gouvernement et, pour chaque message posté, ils gagnent 50 centimes. J'ai un ami qui gagne comme cela quelques centaines de Renminbis par mois.

Qui sont ces gens ? Comment sont-ils recrutés ?

Cela dépend. Mon ami travaille dans des médias traditionnels, il est reporter pour un journal et il complète son salaire en allant poster des commentaires sur les blogs.

D'après mes informations, les autorités recrutent chez les journalistes, mais il y a aussi des personnes employées par le gouvernement à temps plein, des membres du Parti. Par ailleurs, Tsinghua, l'université d'élite chinoise, est un des viviers de recrutement pour ces agents d'information. Les recruteurs s'informent d'abord sur les affinités politiques des étudiants, puisqu'ils cherchent des sympathisants du Parti. Et ils choisissent surtout des hommes, puisqu'il faut naviguer sur des sites pornographiques. Une amie à moi s'est vu refuser sa candidature pour cette raison.

Comment ça marche ?

Les "50 cents people" ne censurent pas directement. Ils essaient d'atténuer les opinions exprimées dans les billets de blogs. Ils s'insèrent dans la discussion en prônant l'avis du gouvernement. En fait, ils façonnent l'opinion publique.

Si j'écris Free Tibet, ils m'attaqueront moi personnellement ?

Certains pourraient t'insulter. Mais il y a différents moyens. »

Une stratégie de riposte qui permet au commanditaire de coordonner des réseaux immergés dans les sites d'influence pour intervenir dès l'apparition d'un mouvement contestataire.

« À la différence du dispositif antirumeur du candidat Obama qui repose sur la transparence et vise à canaliser les rumeurs sur Fight the Smears, l'efficacité des "50 cents people" s'appuie sur une démarche d'infiltration anonyme permettant de participer aux conversations de manière diffuse. On se souvient également de la grande mobilisation chinoise prétendument spontanée pour appeler au boycott de Carrefour et dont la réussite n'est certainement pas étrangère à la présence des "50 cents people" sur le Web communautaire chinois. » [8].

Il serait faux de croire que la réputation numérique dont jouissent une entreprise, une marque, une personne ou un État soit le résultat parfaitement neutre de l'opinion libre des internautes. Nous aurions

tout aussi tort d'imaginer que ces agents d'influence exercent seulement en Chine : chaque État dispose de ses propres propagandistes ou leaders d'opinion qui interviennent régulièrement pour orienter les discussions lorsqu'elles prennent une tournure qui dérange.

Certaines entreprises proposent ainsi des tribunes, des droits de réponse, des commentaires en réaction à des articles, interventions dans les forums, blogs, sites participatifs en adaptant le vocabulaire, le format et le style aux médias récepteurs. Certes, les données semblent truquées et l'on pourrait douter du bien-fondé de ce type d'offres ; cependant, elles sont absolument indispensables lorsqu'une organisation est victime d'une attaque par ces mêmes moyens d'un concurrent malintentionné. Si cela avait été le cas pour Belkin, le fabricant aurait dû faire appel à une entreprise spécialisée dans cette forme de communication plutôt que de tenter de s'en sortir tout seul. En tout état de cause, rien ne peut ni ne doit s'improviser. Si l'on imagine que les critiques étaient fondées et que le produit présentait des défaillances ou des performances en dessous de celles que la clientèle attendait, Belkin aurait dû alors choisir une réponse claire et proposer des solutions aux consommateurs mécontents ; toutefois, cette honnêteté devait nécessairement passer entre les mains de professionnels dont l'objectif aurait été de l'assister dans la rédaction d'un texte adéquat.

L'e-réputation, phénomène récent, représente une arme redoutable pour les individus et les organisations. Nous faisons face à un basculement entre deux cultures : l'ancienne est convaincue qu'il n'y a jamais de fumée sans feu, et la nouvelle a acquis la capacité du buzz Internet de nuire comme d'encenser. Actuellement, il faut encore jongler avec les deux et surtout s'habituer à veiller sur sa propre image dans le monde virtuel dont les impacts sont, eux, bien réels.

Notes

[1] www.cybion.fr
[2] Swot : synthèse d'une analyse marketing identifiant les forces (strengths), les faiblesses (weaknesses), les opportunités (opportunities) et les menaces (threats).
[3] En 2008, 50 millions de blogs étaient recensés dans le monde, dont 3,2 millions de blogs français.
[4] Source : « Digimind White Paper: Réputation Internet », www.digimind.com
[5] Source : Erwan Seznec , *in ufc Que choisir*, janvier 2009.
[6] Source : www.thomascrampton.com
[7] Source : http://www.paperblog.fr/933996
[8] Source : http://www.paperblog.fr/users/opinionwatch/

Chapitre IV

L’accès à, et la manipulation de l’information

L’arrivée du haut débit accompagnée de la baisse des coûts de connexion et du matériel permit aux organisations et aux particuliers de publier sans limite sur le Net. Enfin, il devenait possible de trouver des informations en ligne, ce qui permettait d’élargir le champ des recherches grâce à ce qui allait devenir la plus grande bibliothèque du monde. Inévitablement, cette capacité de recherche et de publication de données ne pouvait qu’aboutir à de nombreuses dérives dont nous commençons seulement à mesurer les effets collatéraux.

I. Les manipulations boursières, les fausses rumeurs

La rumeur, quelle que soit son origine (volontaire ou non), est une nouvelle qui s’infiltre et se répand inexorablement sur la Toile, laissant pour toujours ses traces venimeuses. Ce qui n’a donc jamais existé devient réalité par un phénomène d’amplification démultiplié par le plus puissant des médias.

Les victimes sont alors confrontées à une situation dans laquelle elles devront gérer une réponse appropriée en prenant le risque de voir la rumeur se confirmer au moindre faux pas de leur part. Pour certaines entreprises cotées en Bourse, l’impact peut avoir des effets dévastateurs. À la fin des années 1990, lorsque les premiers sites boursiers sont apparus sur le Web, certains traders lançaient des rumeurs pour influencer les cours. Des entreprises en ont fait de même pour atteindre leurs concurrents et quelques salariés en ont également profité pour discréditer leur employeur. Des travaux effectués en 2005 [1] par le professeur Thorsten Hens, à l’institut de recherche économique de l’université de Zurich, démontrent que les traders utilisent de vrais réseaux pour répandre informations et rumeurs. Il explique aussi que si un cambiste a vent d’une rumeur, il observe tout d’abord comment le cours de l’action en question évolue. Si celui-ci grimpe, le cambiste doit partir du principe que d’autres ont déjà réagi à la rumeur. Ainsi le cours peut augmenter de la même façon qu’avec une information avérée, cela dépend simplement du fait que d’autres personnes sur le marché ont cru, ou pas, à l’information. La rumeur est alors considérée comme un outil stratégique en intelligence économique pour mettre en œuvre désinformation et contre-information. Dès le début des années 2000, des prestataires ont proposé aux entreprises d’infiltrer les forums et les groupes de discussion sur Internet pour y diffuser des messages sur les produits ou les services de la concurrence dans le but de peser négativement sur l’opinion.

En d’autres circonstances, ces messages servent les fraudeurs. En 2000, la Securities and Exchange Commission (sec) a interpellé 33 fraudeurs, accusés d’avoir répandu de fausses rumeurs et informations sur des sites boursiers, des forums et par e-mails. Ils avaient réussi à gonfler artificiellement la valeur de quelque 70 titres boursiers qui leur auraient permis de réaliser des gains de plus de 10 millions de dollars.

Mais, parfois, les rumeurs sont simplement dues aux erreurs de leurs émetteurs qui, par paresse ou étourderie, n’ont pas pris le soin de vérifier les informations qu’ils publient. En août 2008, la compagnie

United Airlines a vu son titre passer de 12,45 $ à 3 $ en à peine plus d'une heure. Un journaliste effectue une recherche sur l'entreprise sur Google et découvre une dépêche du *Chicago Tribune* annonçant la prochaine faillite de la compagnie aérienne. Il s'empresse de rédiger un article et avertit Bloomberg et les places boursières. L'impact est dévastateur : en l'espace de 10 minutes, 24 millions de titres changent de mains. Seulement voilà, l'information datait de plus de six ans (10 décembre 2002) et ual avait, depuis, surmonté la crise [2].

En octobre 2008, ce fut le tour d'Apple. La plate-forme de journalisme collaboratif iReport de cnn publiait un billet posté par un utilisateur anonyme qui affirmait : « Steve Jobs, le président d'Apple, a été transporté aux urgences à la suite d'une grave crise cardiaque. Une source proche m'indique que des infirmiers ont été appelés en urgence quand Steve Jobs s'est plaint de douleurs à l'abdomen et de problèmes respiratoires. Ma source préfère rester anonyme mais est sérieuse. Pas de nouvelles sur ce sujet, nulle part sur le Net pour l'instant, et pas de nouvelle information, si vous en savez plus, tenez-nous au courant [3]. » Malgré les démentis de la firme américaine, le cours a brutalement chuté et atteint son niveau le plus bas depuis plus d'un an pour finir la journée avec une baisse de 3 %. Le site a retiré le post* dans la matinée, mais le mal était déjà fait. La sec (commission de surveillance de la Bourse aux États-Unis) a ouvert une enquête pour déterminer s'il y a eu tentative de manipuler les cours.

Aujourd'hui, les entreprises ont tout intérêt à détacher un salarié (sous anonymat) sur les forums et autres sites d'échange d'information pour tenter d'y devenir un leader d'opinion et pouvoir réagir face à d'éventuelles attaques intentionnelles ou non en publiant des arguments crédibles.

Sur les forums, les modérateurs tracent les utilisateurs pour tenter de définir si un post a été créé dans un but commercial par une entreprise ou de dénigrement par un salarié ou un concurrent. La recherche s'oriente sur le pseudo de l'internaute qui a tendance à utiliser le même surnom sur plusieurs sites. La lecture de quelques-uns de ses posts permet dès lors de s'assurer de la sincérité de l'auteur.

Les sites de grande audience, quant à eux, emploient des personnes pour répondre à des alertes d'internautes ou réagir de leur propre initiative en cas de doute. Les forums féminins sont, pour leur part, fréquemment ciblés par des entreprises qui souhaitent profiter de ces lieux d'échange et de partage pour accroître leur notoriété en annonçant la mise en ligne d'une nouvelle boutique ou la parution prochaine d'un nouveau produit. Il convient donc d'être prudent au risque d'éliminer des utilisateurs sincères mais dont le style rédactionnel hésitant est parfois trompeur.

II. L'espionnage industriel

Les organisations sont vulnérables sur la Toile lorsqu'elles deviennent la cible d'attaquants déterminés à obtenir des données sensibles par tous les moyens. Internet recèle des informations *a priori* anodines mais dont l'utilisation par des personnes malintentionnées peut se révéler dangereuse si leur collecte est destinée à une attaque du système d'information.

Ces données, autrefois exploitées quasi exclusivement par les pirates informatiques, sont aujourd'hui prisées par les officines privées (dites d'intelligence économique), les services de renseignements, tout comme les veilleurs d'entreprises publiques ou privées.

Les informations recherchées sont le plus souvent axées sur les salariés : le Web 2.0 offre de multiples ressources exploitables qui vont pouvoir étayer le dossier d'un candidat à un poste stratégique, ou bien

celui d'un précieux collaborateur qu'une entreprise cherchera à débaucher ou, si nécessaire, en cas de refus, dont elle tentera d'entacher la réputation. Ces données sont les premières exploitées pour le social engineering, technique de manipulation qui vise à obtenir des informations confidentielles auprès de ressources à accès restreint en abusant de la crédulité ou de l'ignorance de personnes internes à l'organisation. Couramment utilisée en informatique pour obtenir les mots de passe, cette technique est aujourd'hui la première approche réalisée par tous ceux qui cherchent à établir un contact direct ou non avec une future victime. Elle permet de connaître les hobbies des cibles et certains éléments clés de leur personnalité qui faciliteront le premier contact.

Les informations sur les sites géographiques sont parfois sensibles et les applications disponibles en ligne posent quelquefois des problèmes inextricables. Le Pentagone a ainsi interdit à Google Earth de publier les photographies de l'intérieur des bases américaines, pour des raisons évidentes de secret défense. Selon le même principe, les détails de certains sites français ont également été masqués sur Géoportail, l'équivalent français de Google Earth.

Reste que certaines bases françaises, comme celle de l'île Longue (base des sous-marins nucléaires), sont floutées sur Géoportail mais parfaitement visibles sur Google Earth ! Cette différence de traitement montre que la protection de l'information sensible ne dépend pas uniquement du risque de sanction pénale, qui n'existe véritablement qu'en cas d'information classifiée, mais surtout de la propre éthique des diffuseurs.

III. Les dérives criminelles (hackers et mafias)

Les dérives sur Internet sont nombreuses et l'information devient stratégique lorsqu'il s'agit de démanteler des réseaux criminels. Dans l'exemple suivant, une organisation criminelle a pu être interpellée grâce à une veille sur le produit fabriqué par l'entreprise.

En 2006, un laboratoire pharmaceutique français a pu découvrir que des places de marché virtuelles, situées en Chine, proposaient la production de son médicament un an avant son autorisation de mise en vente sur le marché. Des organisations criminelles avaient mis en place des réseaux de production principalement situés en Chine et offraient des packages complets aux acheteurs. Ces packages comprenaient des circuits de distribution, *via* des pharmacies en ligne établies dans des zones offshore et les spammeurs inondaient les boîtes électroniques pour assurer les ventes. Cette affaire a été découverte lors d'une recherche sur Internet, sur les sites commerciaux qui annonçaient la vente du médicament. Les organisations criminelles utilisent fréquemment les liens publicitaires pour s'assurer un taux de connexion élevé et surtout pour obtenir un affichage sur les premières pages de résultats des moteurs de recherche.

Il arrive fréquemment que la négligence soit à l'origine de l'infraction. En décembre 2008, 588 start-up ont vu l'intégralité de leurs dossiers soigneusement préparés pour lever des fonds (business plans, informations financières, composition du board, biographies complètes et démos des produits développés) publiée sur Internet. À l'origine de la fuite, la mise en ligne, par le fournisseur d'accès Meteora Technologies Group, de la base de données sql, dont le contenu s'est fait immédiatement aspirer et indexer par Google. Les moteurs de recherche peuvent sembler parfois intrusifs, d'où l'importance de rédiger un cahier des charges précis de ce qui doit être mis en ligne ou pas de manière à éventuellement protéger certaines données par un cryptage adapté à leur niveau de confidentialité.

Notes

[1] Andreas Merz, « Le commerce des rumeurs ».
[2] James Erik Abels, « Inside the ual Story Debacle ».
[3] Voir http://fr.techcrunch.com/2008/10/04/du-journalisme-citoyenqui-fait-trembler-laction-dapple/ et http://www.businessinsider.com/2008/10/apple-s-steve-jobs-rushed-to-er-after-heart-attack-says-cnn-citizen-journalist.

Chapitre V

Le Web 2.0 et les réseaux sociaux, l'envers du décor

L'explosion du Web 2.0, qui offre aux internautes un contenu généré directement par des utilisateurs inconnus, représente une menace considérable pour l'entreprise. Le pire du cybercrime pourrait bien venir de l'exploitation de failles liées aux sites communautaires et aux applications dynamiques en ligne. Elles seules savent répondre aux besoins de communication et de collaboration des entreprises en temps réel.

Ainsi, faisant fi de toutes les procédures de sécurité, les salariés utilisent désormais des applications Web hébergées hors du système d'information de l'entreprise (comme celles offertes par Google) pour accélérer le flux d'information et de productivité des collaborateurs.

L'engouement général pour les sites dynamiques a modifié le comportement des employés, sédentaires et nomades, qui manipulent sans contrôle ni modération les données les plus sensibles de leurs sociétés au travers d'applications en ligne, à des fins légitimes ou non. Les administrateurs système, n'étant pas formés (en termes de sécurité) à ces nouvelles technologies, se cantonnent à sécuriser le réseau par un filtrage des pièces jointes des e-mails, des connexions à des sites à caractère pornographique ou à l'analyse de signatures issues de codes malveillants recensés par les antivirus.

I. Collecte d'informations militaires

Les réseaux sociaux, quant à eux, regorgent d'informations précieuses. Les premières cibles sont les militaires dont l'isolement et l'éloignement les incitent à communiquer à leurs proches leurs états d'âme : bien souvent, ils se laissent aller à raconter des anecdotes vécues sur les théâtres d'opérations, à publier des photos d'eux et de leurs camarades et par conséquent à divulguer des informations pouvant être utilisées par les forces ennemies.

Ce fut notamment le cas pour les soldats britanniques en Grande-Bretagne, ainsi que les Canadiens et les Américains, souvent victimes de ces publications en Irak. Les organisations proches du mouvement Al-Qaida collectaient ces données pour localiser les militaires et perpétrer des attentats. Marine Chatrenet a réalisé un rapport en 2008 pour le Centre d'études en sciences sociales de la défense. Elle y dresse une typologie de ces journaux de bord qui, nourris de photos et de vidéos, livrent des détails sur les camps, les manœuvres et les interventions militaires. « Si les médias en opex (opérations extérieures) sont encadrés et ne peuvent pas tout filmer, les militaires, avec de simples appareils, ont une marge de manœuvre supérieure et surtout l'exclusivité de certaines images », grâce aux téléphones portables dont la discrétion et la capacité de stockage augmentent et qui permettent d'illustrer des propos édités en ligne, souvent dans le plus parfait anonymat et sans aucun contrôle possible de l'armée. Début 2008, l'armée canadienne demandait à ses soldats de ne publier aucune photo ni information personnelle sur les sites de réseaux sociaux en raison des risques avérés d'attentats.

II. Facebook, MySpace, une source pour les services de renseignements comme pour les malfaiteurs

Désormais, les services de renseignements recrutent sur les réseaux sociaux et plus particulièrement sur Facebook. Le mi6, service du renseignement extérieur britannique, publie des annonces pour embaucher du personnel. D'autres services utilisent ce moyen pour cibler des profils intéressants, car les réseaux sociaux se révèlent d'excellents outils de renseignement, permettant à ces organisations d'y glaner de précieuses informations. Ils peuvent aussi en être victimes comme sir John Sawer, le nouveau patron des services secrets britanniques, le mi6. Son épouse, Shelley, a publié en juin 2009, sans aucune forme de protection, sur le site communautaire Facebook des photographies de son mari en maillot de bain, de ses enfants et de leur cercle d'amis. Les adresses des logements du couple et de leurs enfants étaient diffusées, tout comme les messages de félicitations des amis diplomates et acteurs quant à la récente nomination du futur directeur du Secret Intelligence Service (sis). L'information révélée par le « Mail on Sunday » a aussitôt été retirée du réseau social, mais le tabloïd a conservé sur sa page Web des photos de la famille de sir John Sawer [1].

La police de l'Office central de lutte contre la cri-minalité liée aux technologies de l'information et de la communication (oclctic), quant à elle, se sert de crawlers, des outils destinés à sonder les sites Internet, pour traquer les délinquants ou dans le cadre d'enquêtes sur des disparitions de mineurs, d'incitation à la haine raciale, de diffamation comme dans la lutte contre la pédophilie et la pédopornographie. Bien souvent, les enquêtes de voisinage ne sont plus nécessaires puisqu'à présent la plupart des fréquentations des suspects sont publiées sur la page du profil.

Entre mai 2008 et janvier 2009, Facebook a exclu 5 585 membres actifs identifiés comme criminels sexuels condamnés. MySpace a, de son côté, supprimé les comptes de 90 000 délinquants sexuels en deux ans. Les jeunes sont facilement piégés par ces nouvelles formes de communication parce qu'ils ont du mal à imaginer que les prédateurs rôdent et savent prendre le profil d'une personne attrayante, de leur âge. Une adolescente de 16 ans s'est fait piéger sur Internet par un homme de 43 ans qui prétendait en avoir 18. Après de longs échanges en ligne, toute disposée à croire en une correspondance amoureuse avec le prince charmant, la jeune fille s'est décidée à lui rendre visite en province, sans, bien évidemment, prévenir ses parents. L'homme qui l'attendait a prétexté venir de la part du jeune homme souffrant, l'a emmenée à l'hôtel et a abusé d'elle pendant trois jours. En septembre 2009, c'est un homme de 24 ans qui est soupçonné d'avoir violé une jeune fille de 16 ans rencontrée par l'intermédiaire du site Facebook. Après avoir pris contact avec la victime par Internet en se présentant comme un photographe, l'homme est allé la chercher à son domicile de Guyancourt. Il a emmené l'adolescente dans un studio pour une séance de poses. Là, il lui aurait arraché ses vêtements avant de la violer avec un préservatif et de la ramener chez elle [2].

Les organisations criminelles spécialisées dans le trafic d'êtres humains utilisent ces réseaux pour faciliter le recrutement de jeunes femmes (souvent mineures) et fournir des services sexuels en ligne. Les profils piégés présentent de jeunes gens au style vestimentaire et musical branché, hip-hop, qui font de fausses promesses de carrières de mannequin à l'étranger et qui se terminent par l'exploitation des victimes dans l'industrie du sexe.

Fin 2008, Facebook a retiré sept pages de groupes utilisées par des néonazis italiens – une intervention effectuée à la requête de parlementaires européens dénonçant ces communautés qui auraient appelé à la

violence envers les Roms.

En août 2009, la compagnie d'assurances Legal &General a informé ses clients d'une hausse prochaine des primes pour les utilisateurs des réseaux sociaux comme Facebook ou Twitter. Selon le rapport commandé à la société The Digital Crime, des millions d'internautes confieraient les moindres détails de leur vie privée à ces sites de grande audience, augmentant de manière significative le risque de cambriolage de leur résidence lors des départs en vacances. Les utilisateurs publient les photos de leur maison, de leur salon, des réceptions qu'ils y donnent, offrant aux yeux des malfaiteurs la possibilité d'effectuer des repérages précis [3].

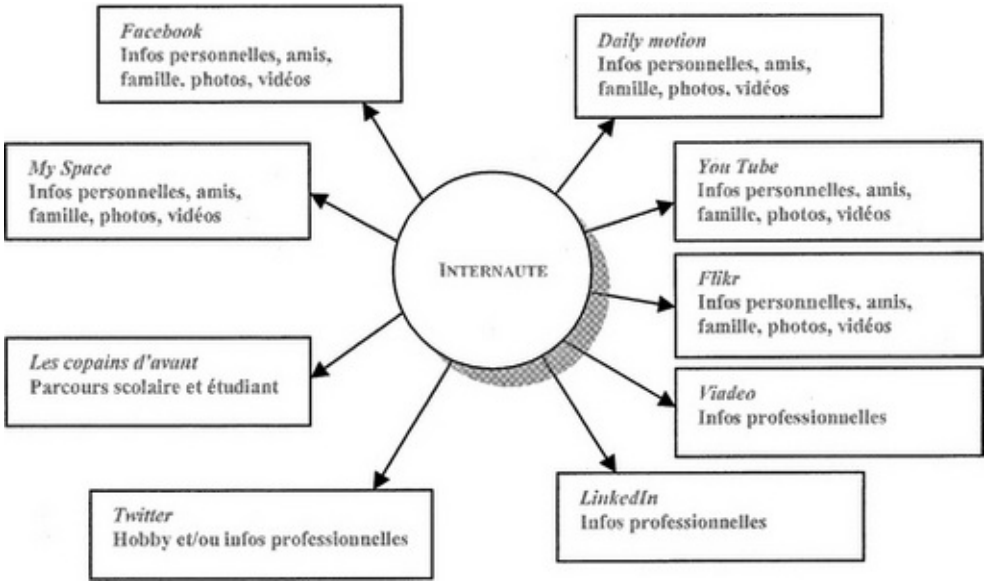
Parfois, la consultation de certains profils s'avère utile. Ainsi, un agent pénitentiaire de la prison de Leicester, en Grande-Bretagne, soupçonné d'introduire des produits illicites pour les détenus, a été licencié après que l'on a découvert que treize de ses amis sur Facebook étaient des criminels. Le gardien de prison avait des relations téléphoniques plus ou moins proches avec sept d'entre eux incarcérés pour divers trafics de drogue, violences, vols et fraudes.

Aujourd'hui, toute personne (victime, criminelle, ou témoin) mêlée à un événement couvert par les médias est immédiatement traquée sur les réseaux sociaux, ses photos sont publiées et tous ses contacts exploités dans le but d'obtenir des témoignages, des commentaires ou pour faire part de son opinion *via* des messages de soutien, de haine ou de menace.

III. Portrait d'un inconnu : la fin de la vie privée

Un internaute, pris au hasard, a eu la désagréable surprise de découvrir sa vie privée détaillée dans un portrait publié par un journaliste du bimestriel *Le Tigre*, le 14 janvier 2009. Ce document a été réalisé à partir d'un concept créé par l'auteur de l'article, selon lequel il s'agissait de rédiger le portrait d'un anonyme grâce à toutes les traces qu'il a laissées, volontairement ou non, sur Internet, sur des sites comme Facebook, Flickr, Les copains d'avant, ou YouTube. Le journaliste réunit deux pleines pages d'informations sur la vie de ce jeune homme d'une trentaine d'années, habitant à Saint-Herblain, et dessinateur dans un cabinet d'architectes. Il est parvenu à reconstituer son parcours professionnel, amoureux et musical en récupérant quelques-unes des 17 000 photos de ses voyages (postées en moins de deux ans), des détails sur sa vie intime, ses passions, les noms de ses ami(e)s et même son numéro de portable.

Une démonstration éclatante de la gestion de l'identité numérique parce que l'on ne fait pas vraiment attention aux informations personnelles que l'on publie au fil du temps sur Internet et qui, lorsqu'elles sont synthétisées, prennent soudain un relief inquiétant. Selon la manière dont sera rédigé un article, la même information pourra être interprétée différemment par le lecteur. Dans ce cas, l'internaute dont le portrait a été publié a plutôt mal vécu cette expérience qu'il a ressentie comme une véritable intrusion, pourtant tous ces détails provenaient exclusivement de sources publiques. Ce qui semble une évidence ne fait pourtant pas partie des réflexes acquis aujourd'hui, chacun trouve logique de consulter les informations disponibles en ligne pour se renseigner sur une personne que l'on vient de rencontrer ou qui postule pour un emploi. Cependant, peu de gens ont pris conscience qu'ils ne devaient pas non plus publier de détails sur leur vie privée comme sur leur vie professionnelle.



IV. Usurpation d'identité

L'usurpation d'identité numérique est surtout connue à travers les attaques de phishing* qui permettent aux malfaiteurs de récupérer les coordonnées précises de leurs victimes à travers l'emploi de sites de leurre comme celui de leur banque ou des sites commerciaux sur lesquels il est fréquent de devoir communiquer ses informations personnelles. Aujourd'hui, elle évolue et se présente différemment au travers des sites de réseaux sociaux. Contrairement au phishing, ces usurpations n'ont pas pour objectif de s'en prendre aux actifs bancaires des personnes ciblées, il s'agit plutôt, dans ce cas, de nuisance collective (massive) déclinée en deux versions.

La première version concerne des personnalités plus ou moins célèbres, qui n'ont pas créé de profil sur les réseaux sociaux. Une personne malveillante créera alors quelques pages sur Facebook ou MySpace et s'appliquera à décrire autant de détails réels qu'imaginaires pour semer la confusion chez les internautes. Tous ceux qui entreront en contact avec le faux profil pour devenir des amis seront acceptés, puis petit à petit, il y aura des inscriptions à des groupes souvent liés à des sujets à caractère pornographique de préférence outranciers, de fausses photos intimes seront publiées, etc.

Une animatrice de radio canadienne en a fait la douloureuse expérience début 2009. Cette jeune femme a pris connaissance de l'existence de son compte sur Facebook par des amis qui affirmaient communiquer avec elle *via* le site, alors qu'elle n'y avait jamais pointé le bout de sa souris. Après vérification, le profil affichait plus de 1 000 amis Facebook, acceptait tout le monde très rapidement et devenait membre de groupes surprenants.

Pour plus de crédibilité, elle était membre de groupes tels que : enfants malades, cancer du sein, etc. Mais là où la fausse animatrice a commencé à aller un peu trop loin, c'est quand elle s'est mise à faire du recrutement pour un ami photographe faisant de la photo nue et qu'elle s'est jointe à un groupe faisant la promotion de certaines pratiques sexuelles. Une attitude totalement impossible pour ceux qui la connaissent réellement.

Ce genre de pages attire immanquablement les foudres des internautes qui imaginent soudain que les personnes qu'ils apprécient (célèbres ou non), sont en fait des êtres pervers ou des malfaisants et qu'ils ont été dupés : cela engendre des commentaires extrêmement désagréables de leur part à l'encontre de la victime et parfois même des actes de violence physique.

Les profils sont généralement supprimés mais cela prend du temps et demande des démarches fastidieuses. Les victimes doivent justifier leur identité et vivent, à juste titre, l'usurpation comme un viol de leur intimité.

La deuxième version, assez nauséabonde, concerne ceux qui se font passer pour des criminels comme Jacques Mesrine – qui totalise plus de 160 500 fans sur sa page –, ou Marc Dutroux qui, après de nombreuses plaintes, n'en a plus que 82, ou encore les Michel Fourniret, Francis Heaulme, Guy Georges ou Émile Louis qui bénéficient de quelques dizaines de pages chacun avec, sur chaque profil, plusieurs centaines d'amis. Dans le même registre, on trouve aussi Joseph Fritzl, cet Autrichien qui a violé et séquestré sa fille pendant vingt-quatre ans dans sa cave. Mentionnons au passage que ces mêmes criminels sont au cœur de nombreux groupes dont les thèmes renvoient à leurs activités délictueuses comme celle intitulée « Comment aménager sa cave by Joseph Fritzl ».

Les profils se créent en quelques minutes avec des e-mails conçus pour la circonstance sur des webmails. Mais si les créateurs de ces pages sont le plus souvent anonymes, les membres ou les amis qui s'y inscrivent, eux, ne le sont pas et ils laissent sur le Net des traces indélébiles. De tels penchants peuvent être fatals pour une carrière professionnelle : être le fan d'un assassin et de son mode opératoire n'est pas avantageux « sur un cv virtuel ».

Finalement, le problème se pose de savoir quelle est l'attitude à observer face aux réseaux sociaux : prendre le risque d'être piraté ou infecté par des virus, ou encore celui de subir une usurpation d'identité avec toutes les conséquences que cela peut représenter, surtout lorsque l'on sait aussi que certains criminels demandent à des subalternes de consulter les photos des profils enregistrés pour y repérer ceux qui présenteraient une ressemblance physique dans le but d'usurper leur identité.

V. Les réseaux sociaux et l'entreprise

Les organisations sont face à un véritable dilemme vis-à-vis des réseaux sociaux. Des études quant aux effets néfastes ou positifs de leur accès depuis les postes de travail sont régulièrement publiées et tout aussi régulièrement contradictoires. En termes de sécurité, leur utilisation serait totalement à proscrire. Mettre son profil sur de tels sites expose l'internaute à de multiples tracas [4], depuis l'infection virale (au 15 mai 2009, 56 variantes de vers ont attaqué Facebook [5]), jusqu'au phishing, au vol d'identité, etc. Les entreprises s'inquiètent de ces sources de distractions autant que du risque de fuite d'informations sensibles, sans compter les tentatives de sollicitation du personnel par la concurrence, car les réseaux sociaux sont la première cible des chasseurs de têtes. Pour les employés, les avantages sont énormes, mais l'entreprise doit être consciente que chaque personne publie à la fois son savoir-faire, ses contacts et ses idées sur une plate-forme publique, donc accessible à n'importe qui. Elles doivent en conséquence, s'assurer d'une forme d'éthique sur ce qu'il est possible d'écrire ou non.

Le Trade-union Congress britannique a opté pour un code de bonne conduite en éditant une charte d'utilisation des réseaux sociaux en entreprise. Il s'agit là de rassurer les organisations hésitantes en proposant un encadrement pour une utilisation responsable et raisonnable de la part des collaborateurs plutôt que de leur en interdire l'accès. Un compromis intelligent dans la mesure où les salariés frustrés iront créer leur profil et alimenter leurs réseaux en dehors de l'entreprise.

Selon une étude de Gartner, les organisations auraient tout intérêt à opter pour des sites de réseaux sociaux populaires (Facebook, Del.icio.us, etc.), plutôt que de créer des communautés en ligne dont le

nombre de participants risque d'être trop faible pour y apporter une réelle valeur ajoutée. Le cabinet Deloitte a déclaré que 35 % des communautés dédiées à la promotion commerciale comptaient moins de 100 membres et que moins du 25 % en avaient plus de 1 000. Alors que sur Facebook, le nombre d'utilisateurs atteint aujourd'hui 250 millions.

Pourtant, les entreprises adoptent les réseaux sociaux : entre fin 2007 et début 2008, ils représentent, selon International Data Corporation (idc), un marché qui pourrait atteindre 2 milliards de dollars en 2012. Ils ont incontestablement une importance capitale pour les organisations. Dell s'est, par exemple, basé sur son site ideastorm.org pour déterminer la demande concernant les pc Linux ; la réaction des membres de ce réseau social a permis au fabricant d'ordinateurs de lancer une production de machines équipées du système d'exploitation Open Source. Les communautés rendent les salariés plus efficaces car elles sont sources d'idées et de créativité, elles permettent d'exploiter et de maîtriser le savoir collectif des employés, des clients et des fournisseurs et offrent davantage de solutions aux problèmes posés. Les réseaux sociaux stimulent l'esprit d'équipe, augmentent la connaissance des individus et favorisent la collaboration interne.

Pour Kathryn Everest, consultante senior en management chez ibm Canada, l'adoption de ces technologies ne peut que favoriser l'entreprise, surtout au niveau de l'apprentissage car les mécanismes de collaboration et de partage de l'information au cœur de l'entreprise 2.0 vont renverser totalement les façons d'opérer jusqu'ici. Pour elle, c'est la seule façon d'innover désormais. « Les gens apprennent beaucoup mieux les uns des autres […]. Pour vraiment innover, il faut des équipes multidisciplinaires, soutient K. Everest. Mais le problème reste : comment faire pour rassembler des gens aux profils variés ? C'est là qu'entrent en jeu les réseaux sociaux. Les réseaux sociaux permettent aussi de réintégrer les retraités dans l'entreprise et ainsi de faire profiter les autres de leur expérience, car il y existe nombre de retraités qui ne veulent pas devenir professionnellement inactifs. »

Les réseaux sociaux permettent aussi aux employés à l'étranger d'interagir plus facilement avec leurs collègues demeurés au pays. Et la consultante de conclure : « Il y a une grande valeur dans les interactions qui ont cours dans les réseaux sociaux et ce n'est pas parce qu'on est loin qu'on ne peut pas en profiter [6]. »

Ces communautés en ligne reflètent la philosophie de l'entreprise et permettent aux employés de renforcer son image à l'extérieur, bien que certains employeurs appréhendent une décentralisation de la communication et une perte de leur pouvoir au profit des salariés. Effectivement, les réseaux sociaux remettent en cause la hiérarchie pyramidale et la direction perd une partie parfois significative de son emprise sur ses salariés ; toutefois, utilisées dans de bonnes conditions et avec confiance, les ressources peuvent alors se recentrer et se mobiliser autour de l'objectif commun de l'entreprise.

Notes

[1] Source : « Mail on Sunday ».
[2] http://lci.tf1.fr/france/faits-divers/2009-09/ecroue-pour-le-viol-d-une-ado-rencontree-sur-facebook-4863496.html
[3] Source : www.thetelegraph.co.uk
[4] MySpace, ciblé par des attaques de spams, avait d'ailleurs obtenu la condamnation du spammeur au

paiement d’une amende de plus de 200 millions de dollars. Outre ces menaces, les sites dits Web 2.0, parmi lesquels Twitter, sont parfois aussi la cible d’attaques virales. Deux vers informatiques ont ainsi été récemment propagés sur Twitter. En 2008, le virus Koobface se propageait sur Facebook. La même année, un cheval de Troie visait MySpace.

[5] L’équipe de sécurité de Facebook considère que les campagnes de phishing visent avant tout à récolter des identifiants de connexion afin, dans un second temps, d’envoyer des messages de spam aux utilisateurs du service.

[6] Source : directioninformatique.com et touscomplices.com

Chapitre VI

Le contrôle de la validité de l'information

Dans le foisonnement d'informations disponibles, le plus complexe reste la capacité à pouvoir la valider. Dans certains cas, cela peut se révéler assez difficile ; par exemple, les services de renseignements publient des sites au contenu identique à ceux des terroristes proches de la mouvance Al-Qaida : même arborescence, mêmes discours, photos, vidéos et forums, tout est très convaincant et a pour objectif d'identifier les internautes qui s'y connectent et qui participent activement aux débats. Comme il est impossible de vérifier qui est le réel concepteur du site, il est aussi impossible de contrôler la validité des informations qui y sont diffusées à moins de procéder à des recoupements complexes. Cette difficulté s'accroît sur la Toile parce que les informations sont régulièrement reprises par d'autres sites, voire parfois par la presse papier, par des journalistes ou des webmasters qui vont, au fil de leur navigation, piocher une information et la relayer sans la vérifier. Le Web est alors inondé de fausses informations, chacun étant certain de ses propres sources.

Un jeune étudiant irlandais de 22 ans a démontré, en mars 2009, que de fausses informations publiées sur Internet pouvaient être reprises par la presse sans la moindre vérification des sources. Convaincu de la dépendance des journalistes au Net et du manque de vérification des informations y circulant, Shane Fitzgerald a décidé de tenter une expérience dont les résultats sont aussi édifiants qu'inquiétants. L'étudiant a attendu patiemment qu'un événement sur lequel il pourrait écrire se produise. Quelques jours plus tard, le décès de l'auteur-compositeur Maurice Jarre est annoncé par SkyNews, Shane Fitzgerald a alors imaginé une citation du musicien et l'a publiée sur la page Wikipédia qui lui est consacrée : « On pourrait dire que ma vie elle-même a été une musique de film. La musique était ma vie, la musique m'a donné la vie, et la musique est ce pourquoi je vais rester dans les mémoires longtemps après que je quitterai cette vie. Quand je mourrai, il y aura une dernière valse jouant dans ma tête, que je pourrai seul entendre. »

Au-delà de toute attente, les mots seront repris intégralement dans des journaux aussi prestigieux que *The Guardian* et *The London Independent*. Mais l'affaire ne s'arrête pas là et la fausse citation est également diffusée aux États-Unis, en Australie et en Inde. Rapidement, Shane annoncera avoir inventé le texte publié sur Wikipédia et les journaux corrigeront l'erreur [1]. Cela démontre l'immense crédibilité accordée au site collaboratif bien qu'il ne soit pas rédigé par des experts, ainsi qu'une paresse intellectuelle qui s'installe dans les esprits des internautes noyés dans la masse d'informations et qui préfèrent se fier à des informations approximatives plutôt que d'approfondir leur recherche.

Certaines personnalités au passé peu glorieux arrivent ainsi à modifier leur portrait grâce à l'aide de quelques amis qui publieront des informations largement édulcorées sur Wikipédia et offriront aux lecteurs un nouveau profil nettement plus « politiquement correct ».

Par conséquent, avant d'évaluer le contenu d'un document, il convient de commencer par valider précisément son identification. Une opération assez complexe lorsqu'il s'agit d'information récupérée en ligne, puisque, par son concept même, le monde numérique tend à dissimuler l'origine des écrits comme

de leurs auteurs dans les méandres du Net. Selon Alexandre Serres, dans une étude réalisée pour l'Unité régionale de formation à l'information scientifique et technique de Bretagne et des Pays de la Loire [2], il s'agit tout d'abord de procéder à une identification sur trois parties intimement liées du document.

Le nom de l'organisation, du site producteur ou hébergeur, autrement dit de celui qui détient la responsabilité éditoriale du document. Dans cette première partie, seront identifiés les points suivants : la nationalité du site, sa nature (universitaire, commerciale, politique, éducative, institutionnelle, etc.), son statut (public, officiel, privé), son objet (recherche et publication de travaux, diffusion d'information, d'expression personnelle, de militantisme, de service, etc.). Quels sont le public visé (professionnel, universitaire, communautaire ou tous publics), la date de création du site et celles des mises à jour et enfin sa notoriété qui par ailleurs ne correspond pas forcément au sérieux ni à la crédibilité du contenu. Ensuite une recherche sera effectuée sur l'auteur du document (physique ou moral) :

Est-il présenté par sa réelle identité ou par un pseudo ? Quel est son statut (entrepreneur, ingénieur, universitaire), est-il qualifié et reconnu dans son domaine de compétence (amateur, spécialiste ou expert) par des personnalités ? Son réseau de références (les auteurs, ouvrages et travaux que l'auteur cite dans son document sont d'importants indicateurs) et, le plus important, ce que semblent être ses motivations. L'auteur peut en effet souhaiter partager ses informations, assurer sa propre promotion ou encore faire de la propagande ou du prosélytisme politique, religieux ou social.

Le troisième et dernier plan concerne la nature du document en consultant en tout premier lieu sa date de publication et les éventuelles dates de mise à jour : situer un texte dans le temps contribue fortement à son évaluation, ce qui n'est pas forcément évident sur certains sujets. Si le document n'est pas daté, une recherche sur une partie du texte permet généralement de le retrouver sur d'autres sites qui auront affiché le jour et l'année de publication. Une attention particulière sera portée au support documentaire, qu'il soit un site Internet, un forum, un site de réseau social ou une base de données.

Ces validations, parfois longues et fastidieuses, restent l'unique moyen de s'assurer de détenir une information relativement fiable.

I. L'absence de confidentialité des recherches sur Internet

Les moteurs de recherche aspirent et indexent le Web pour recenser les milliards de pages qui circulent sur la Toile. En fonction de la technologie utilisée, ces moteurs doivent héberger de nombreux serveurs. Google est passé de 100 000 serveurs en 2004 à environ 2 millions en 2008 [3], l'ensemble étant réparti sur 147 sites dans des Google Farms, dont les Datacenters sont de la taille d'un terrain de football. Cette gigantesque infrastructure est basée sur une multitude de petites machines de moins de 1 000 dollars stockées dans des conteneurs pour des raisons de coûts et de redondance maximale. Google a été le premier moteur de recherche à utiliser des cookies (qui expireront en 2038, depuis l'interdiction, aux États-Unis, d'émettre des cookies permanents). Lors de la première connexion sur Google, un cookie est envoyé à l'utilisateur et installe un numéro d'identification unique sur son disque dur. Par la suite, à chaque connexion, le cookie transmet ce numéro d'id qui comporte l'adresse ip de l'internaute, la date et l'heure, la configuration du navigateur. Les requêtes de chaque utilisateur sont donc conservées et archivées sur les serveurs de Google ; à travers elles, le moteur de recherche collecte des informations sur les goûts de l'utilisateur, ses centres d'intérêt et ses croyances, ces différentes données étant susceptibles d'être utilisées par des tiers (des groupes religieux, des publicitaires, ou encore des services de renseignements). En 2007, l'Europe s'inquiétait déjà de la conservation des données par Google, lui

reprochant d'enfreindre des lois communautaires sur la protection de la vie privée en conservant les données relatives aux recherches effectuées par ses utilisateurs sur une période pouvant aller jusqu'à deux ans [4]. Selon Google, cette collecte est effectuée pour mieux connaître les attentes de ses utilisateurs, améliorer les performances de ses services et répondre à des questions de sécurité. Mais plusieurs groupes de défense des libertés individuelles craignent une exploitation commerciale et marketing de ces données.

Ce problème se heurte à celui de la justice dans le cadre de la lutte contre le terrorisme et la pédophilie. Les services de police et les magistrats ont en effet besoin de recueillir des informations concernant les recherches effectuées par certains suspects mis en examen. Cependant, elles peuvent être utilisées à l'encontre de n'importe quel internaute, notamment dans les pays qui opèrent une censure sur les sites et sur les requêtes. En avril 2005, le journaliste Shi Tao a été appréhendé, jugé et condamné à dix ans de prison pour divulgation de secrets d'État à l'étranger. La pièce principale du procès dont il est question était une note du journaliste incriminé à propos d'un communiqué du gouvernement sur des événements de la place Tianan men. Logiquement introuvable, cette note n'aurait pas été démasquée par les autorités chinoises à Hong Kong mais par Yahoo ! qui a transmis les informations présentes sur le compte mail de Shi Tao. Ce procédé légal est aussi une règle pour les fournisseurs d'accès ou toute organisation médiatique sur le territoire chinois.

En février 2006, le département de la Justice des États-Unis a sommé Google de lui ouvrir l'accès à ses serveurs et de lui remettre plus de 1 000 requêtes d'utilisateurs qui seraient analysées par le gouvernement. Yahoo ! et Microsoft ont reçu les mêmes demandes et ont immédiatement obtempéré [5].

II. Recoupement des requêtes

Lorsqu'en août 2006, une équipe d'Aol publie sur le Net [6], les logs* de 657 000 Américains dans le but de favoriser la recherche universitaire, le monde prend soudain conscience des risques d'un croisement des requêtes de chaque utilisateur. Aol ne craignait en rien l'identification des internautes, dans la mesure où les cookies ne sont que des numéros de session aucunement liés au nom de l'utilisateur. L'objectif de cette publication, qui regroupait environ 20 millions de requêtes effectuées pendant les mois de mars, avril et mai 2006, était plutôt l'analyse générale que l'on pouvait en faire : par exemple, 45 % des clics se faisaient sur le premier résultat affiché contre 13 % pour le deuxième ; un pourcentage significatif de requêtes est mal orthographié, et un tiers des recherches faisait l'objet d'une reformulation par les internautes.

Les noms des utilisateurs avaient été soigneusement remplacés par les numéros des cookies pour tenter de protéger l'anonymat des fichiers mais l'éditeur n'avait pas un seul instant envisagé que la compilation des logs d'un même numéro de session puisse permettre d'en retrouver l'émetteur. Des journalistes du *New York Times* se sont amusés à mettre bout à bout les requêtes issues des mêmes numéros de cookies. C'est ainsi que de nombreuses personnes ont pu être précisément identifiées, les logs trahissant l'intimité la plus profonde des internautes.

Chaque personne possède son jardin secret et aucun utilisateur ne tient à ce que ses recherches sur Internet soient publiques ; certaines sont anodines, d'autres beaucoup moins, comme le prouvent celles émises par un Américain habitant la Floride (l'utilisateur 14162375), soupçonnant son épouse d'avoir une relation extra- conjugale. Pendant trois mois, les différentes recherches qui, au fil des jours, démontrent ses états d'âme, sa frustration et son désespoir sont visibles. Sentiments qu'il ne souhaitait

certainement pas partager avec le reste du monde :

Mars 2006
Sauver son mariage
Techniques sexuelles
Stopper son divorce
Arrêter l'alcool
Les symptômes de l'alcoolisme *(recherche effectuée à 10 h du matin)*
Problèmes d'érection
Espionner à distance
Écouter à travers les murs
Enregistrer les conversations dans une voiture
Avril 2006
Espionner sa femme
Prédire mon avenir
Je veux me venger de ma femme
Se venger d'une femme qui triche
Divorce et enfants
Ma femme veut me quitter
Comment retrouver l'amour de ma femme
Besoin d'aide pour récupérer ma femme
Ma femme ne m'aime plus
Faire souffrir ma femme autant que je souffre
Revanche du mari
Comment faire du mal à l'amant de ma femme
Infidélité
Je veux me suicider
Je veux tuer l'amant de ma femme
Enregistrement et surveillance du domicile
Enregistrement audio d'une chambre
La mafia portugaise
Mai 2006
Assurance moto
Vidéosurveillance
Pages blanches

Thelma Arnold a également été identifiée au travers de ses recherches sur le moteur d'Aol. Cette internaute de 62 ans, vivant à Lilburn, dans l'État de Géorgie, est enregistrée sous le numéro 4417749. La simple compilation de ses requêtes a permis de la retrouver. Ses recherches concernaient des informations sur les propriétaires de Lilburn et sur les personnes dont le nom de famille est Arnold, puis sur les chiens qui urinent partout. Quelques questions nous renseignent sur ses centres d'intérêt et sur ses préoccupations, comme « termites », « fournitures scolaires pour les enfants irakiens », « quel est le lieu le plus sûr où vivre », « la meilleure saison pour visiter l'Italie ». Ces informations collectées renseignent sur l'auteur des requêtes, cependant elles peuvent aussi être trompeuses. Les recherches de Thelma Arnold concernaient aussi « les tremblements des mains », « les effets de la nicotine sur le corps », « la bouche sèche » et « la bipolarité », ce qui aurait pu amener à penser que l'internaute présentait un

début de la maladie de Parkinson, qu'elle fumait et qu'elle souffrait de troubles bipolaires. Or, lors de son entretien avec les journalistes, Thelma Arnold a déclaré se renseigner pour ses amis, afin de les aider et de les soutenir, comme dans le cas de l'une d'entre elles qui tentait d'arrêter de fumer sans y parvenir. Il devient dès lors indispensable de ne pas faire d'amalgame inconsidéré des données sous peine d'en tirer des conclusions totalement erronées.

Dans un autre exemple, l'internaute 11110859 a cherché où acheter des « fringues hip-hop », puis une rencontre probable dans sa vie de jeune femme lui fait rechercher des informations sur « la perte de virginité » avant de lancer des requêtes frénétiques trois semaines plus tard sur le risque d'être enceinte. Un peu plus tard, il semble que la relation s'effrite au constat de ses questions telles que : « Comment aimer quelqu'un qui vous maltraite ? », ou : « Que dit Jésus au sujet d'aimer son prochain ? » Elle cherche ensuite à acheter des timbres à l'effigie de Betty Boop, puis contacte la direction de la prison de Rikers Island à New York, avant de demander ce que l'on est autorisé à y apporter.

La solitude de l'internaute dans cet exemple est flagrante : le besoin de trouver des réponses à des questions qu'elle ne peut, de toute évidence, poser ni à sa famille ni à ses amis et qu'elle confie sans pudeur au moteur de recherche. La lecture de ses requêtes procure un malaise et révèle un voyeurisme malsain. Toutefois, dans certains cas, la collecte de données pourrait permettre de retrouver des prédateurs sexuels qui effectuent des recherches poussées sur la pornographie enfantine ou sur les apprentis terroristes qui cherchent à se former sur les sites propagandistes. Ces deux types d'internautes passent l'essentiel de leur temps de connexion à chercher des informations issues de sites, forums ou blogs spécialisés. La compilation de leurs connexions pourrait dévoiler des réseaux qui s'échangent des plates-formes de téléchargement ou des informations liées à leurs activités.

Le site www.aolstalker.com permet toujours de fouiller parmi les 36 389 569 requêtes effectuées entre le 1er mars et le 31 mai 2006. Cette base de données offre deux niveaux de recherche ; le premier, par mots-clés qui indiqueront les utilisateurs ayant fait la même requête, par exemple la recherche sur le mot « election » affichera :

1. User 8041 searched for "*jimmy carter's election*" [!] at 2006-05-12 19:25:49 (found http://www.jimmycarterlibrary.org)
2. User 8041 searched for "*jimmy carter's election*" [!] at 2006-05-12 19:25:49 (found http://www.presidentsusa.net)
3. User 8041 searched for "*jimmy carter's election*" [!] at 2006-05-12 19:25:49 (found http://www.multied.com)
4. User 8041 searched for "*jimmy carter's election*" [!] at 2006-05-12 19:25:49 (found http://www.cartercenter.org)
5. User 8041 searched for "*jimmy carter's election*" [!] at 2006-05-12 19:25:49 (found http://www.ourgeorgiahistory.com)
6. User 113945 searched for "*tennessee election results website*" [!] at 2006-05-02 10:40:17 (found http://www.votetn.com)
7. User 113945 searched for "*williamson county tn election results*" [!] at 2006-05-02 10:42:27
8. User 122581 searched for "*election*" [!] at 2006-05-15 20:39:13
9. User 122581 searched for "*pa general election*" [!] at 2006-05-15 20:42:07
10. User 122581 searched for "*delaware county election day*" [!] at 2006-05-15 20:49:05 (found http://www.co.delaware.pa.us)

« Results for election, showing result 1 to 20 of 2 042 search terms. Search took 0.09 seconds » :

Le deuxième niveau de recherche s'effectue sur un numéro d'utilisateur, par exemple le 8041 (extrait des requêtes) :

Query	Querytime	Click URL
lyrics to les choristes [!]	2006-03-01 18:25:30	http://www.ostlyrics.com
im not a perfect person lyrics [!]	2006-03-06 19:33:33	http://www.lyricsandsongs.com
i found a reason to change who i was lyrics [!]	2006-03-06 19:34:31	
i found a reason to change who i was lyrics [!]	2006-03-06 19:34:36	http://www.lyrics007.com
laws of thermodynamics [!]	2006-03-06 21:29:35	http://www.emc.maricopa.edu
aks [!]	2006-03-07 17:20:29	
walt disney company [!]	2006-03-07 18:10:26	http://corporate.disney.go.com
new york stock exchange [!]	2006-03-07 18:12:30	http://www.nyse.com
new york stock exchange [!]	2006-03-07 18:12:30	http://www.investopedia.com
new york stock exchange [!]	2006-03-07 18:15:51	http://www.nyse.com

« Queries made by #8041 on the Aol search engine » :

Comble de l'ironie, la première recherche effectuée sur Aol est Google, les autres sont très diverses et parfois inquiétantes :

« Comment se suicider »

« Comment tuer ma femme »

« Comment tuer sans se faire prendre »

« Comment tuer son petit ami »

« Je déteste mes enfants »

« Je veux devenir une star du porno »

Etc.

Sans compter les nombreuses demandes qui incluent le nom et/ou l'adresse de l'internaute.

Dès l'apparition des premières compilations diffusées par les internautes, Aol a retiré le document en ligne, mais ce qui est publié sur la Toile y reste pour toujours et les 439 Mo de fichiers compressés (soit environ 2 Go, une fois décompressés) peuvent donc toujours être téléchargés sur les réseaux Peer-2-Peer.

The Guardian s'interroge sur le vrai pouvoir totalitaire qui ne repose pas tant sur la censure (comme dans certains pays où des requêtes sont interdites et filtrées [7]) que sur le fait de laisser toute liberté aux citoyens et de stocker leurs recherches.

Partant du principe que chaque requête est enregistrée, on peut imaginer que l'internaute est entièrement scanné et que ses centres d'intérêt peuvent être utilisés contre lui, mais il est tout aussi simple d'envisager que les réponses à ses requêtes peuvent être « manipulées », c'est-à-dire qu'en fonction du profil d'un utilisateur, il est possible de lui envoyer des pages Web, certes réelles, mais triées et sélectionnées en fonction de l'orientation désirée, ce qui est déjà le cas en Chine qui voit en Internet un outil de propagande pour ses citoyens et de désinformation pour les puissances étrangères. On estime à 10 % le contenu Internet fourni par le gouvernement destiné à réorienter les internautes vers une information considérée comme seule acceptable.

III. Les cookies

Ces petits fichiers textes stockés par le navigateur sur le disque dur de l'internaute servent principalement à enregistrer les préférences du visiteur ou encore son parcours sur le site Internet consulté.

Les concepteurs des sites définissent eux-mêmes les informations qu'ils souhaitent obtenir par les cookies. Exalead, par exemple, conserve (sur le poste de l'utilisateur et non sur ses propres serveurs, sauf en cas d'enregistrement auprès de l'éditeur) les préférences de l'internaute telles que la langue (français, anglais, espagnol, etc.), le nombre de résultats souhaités par page, ses raccourcis, etc. Lorsque l'utilisateur se reconnecte sur le moteur d'Exalead, il renvoie son cookie et les paramètres stockés lui évitent de les définir à nouveau. Dans d'autres cas, les cookies permettent de garder en mémoire les identifiants de connexion pour éviter à l'internaute de saisir à chaque fois son nom et son mot de passe pour se faire reconnaître. Chez Google, les données sont conservées dix-huit mois et indiquent l'ip, la date et l'heure de connexion, ainsi que les requêtes, ce qui, lorsqu'on voit les résultats de trois mois de compilation chez Aol, en dit long sur ce que le géant peut savoir sur les millions d'utilisateurs qui s'y connectent chaque jour.

Mais il est important de comprendre que les informations conservées sont d'abord liées à un numéro de session du cookie et que, si ce dernier est effacé (dans le répertoire qui stocke les fichiers temporaires d'Internet), la prochaine connexion ne sera pas liée au même utilisateur, car le moteur de recherche enverra alors un nouveau cookie. De plus, dans la grande majorité des cas, les internautes bénéficient d'une adresse ip dynamique attribuée par le fournisseur d'accès, et, si l'ordinateur est éteint, il se verra attribuer, lors de la prochaine connexion sur Internet, une nouvelle adresse ip ; si, au contraire, l'utilisateur a une ip fixe, alors, quel que soit le numéro de session du cookie, l'ip identifiera toujours le même internaute. Éviter d'avoir ses requêtes recoupées est donc assez simple, mais peu d'utilisateurs savent comment nettoyer les fichiers temporaires sur leur ordinateur et, parmi ceux qui le savent, peu d'entre eux prennent la peine de le faire.

Quel est donc l'intérêt pour les éditeurs de consulter les logs de connexion ? Il s'agit tout d'abord d'améliorer le temps de recherche des requêtes ; en effet, celles qui sont souvent sollicitées par les internautes seront conservées sur les serveurs des éditeurs pour éviter un engorgement et un ralentissement de la bande passante. Les réponses parviennent donc plus rapidement aux navigateurs. Les fautes d'orthographe sont aussi intégrées et permettent de faire comprendre au moteur que « Sarkosi » est probablement lié à « Sarkozy ».

Les numéros de session permettent aussi de connaître le taux de fidélisation des internautes : reviennent-ils régulièrement ou sont-ils des visiteurs uniques ? Les moteurs s'intéressent au succès de leur outil par pays et les résultats permettent d'améliorer les fonctionnalités selon les spécificités liées aux différentes cultures des internautes.

Le Web est donc un vecteur de recherche et développement, un laboratoire géant qui annonce les tendances et permet de tester de nouveaux services dont le succès ne peut être déterminé que par les utilisateurs. La politique est totalement différente selon les éditeurs : Google tient à apposer sa marque et loue son moteur à des marques qui doivent afficher « powered by Google » ; Exalead préfère vendre son produit en marque blanche : la souplesse de sa technologie lui donne la possibilité de modifier les paramètres d'affichage et de résultat selon la volonté de ses clients. Certains préfèrent que les requêtes privilégient les blogs, d'autres optent pour les sites commerciaux ou institutionnels, les résultats apparaîtront alors dans l'ordre préalablement défini. Imaginons, par exemple, un portail sur les voitures : celui-ci préférera que les requêtes aboutissent en première partie sur des sites commerciaux et sur des sites d'annonces plutôt que sur des blogs et des forums. Une offre qui permet aux éditeurs de sites de définir un style rédactionnel adapté à une audience ciblée plutôt que de fournir une information en vrac dont les deux tiers au moins ne présenteront pas le moindre intérêt pour le visiteur.

IV. Anonymat et préservation de la vie privée

Il existe deux moyens de rester anonyme sur le Net, le premier étant d'utiliser un proxy* [8] qui, de par sa fonction d'intermédiaire, se connectera à Internet pour le compte de l'utilisateur et indiquera au site visité son adresse ip à la place de celle de l'internaute. Plus il y aura de proxies entre l'utilisateur et Internet, plus l'anonymat sera garanti. Des moteurs de recherche, sensibilisés par le besoin de préservation de la vie privée, ont décidé de ne pas enregistrer les adresses ip des internautes. Par exemple, le métamoteur hollandais Ixquick assure effacer à la fois l'adresse ip et le cookie d'identification de l'utilisateur. Toute sa communication, fondée sur la garantie d'anonymat de ses utilisateurs, permet de comprendre rapidement les capacités réelles des moteurs de recherche à conserver et à exploiter les informations collectées lors d'une connexion.

Notes

[1] Source : http://www.msnbc.msn.com
[2] https://www.uhb.fr/urfist/
[3] Source : *The Economist*.
[4] Source : Reuters.
[5] Source : DailyTech.com
[6] Sources : *The Guardian*, *The New York Times*, août 2006.

[7] Disposant d’un arsenal technologique de pointe, le système de censure chinois est en mesure de filtrer les sites Web ou leur contenu à partir de mots-clés. Il exerce un contrôle total des échanges de courriels et peut accéder à tous les ordinateurs personnels ou non. Les fournisseurs d’accès Internet (fai) sont tenus de conserver à la disposition du gouvernement les coordonnées de tous les utilisateurs inscrits, ils doivent aussi signaler immédiatement tout débordement. Les mots « démocratie » et « liberté » sont interdits sur les moteurs de recherche.
[8] Tor est l’un des plus célèbres proxies utilisés dans le monde.

Chapitre VII

Organisation de l'information

La capacité de l'humain à gérer les informations est arrivée à saturation devant le vertige que procure l'infinie profusion de documents circulant à la fois sur le Net et sur les ordinateurs de l'entreprise. Chacun s'emploie à collecter des données et à les classer du mieux possible pour les retrouver en temps voulu. Mais le plus organisé des documentalistes ne saurait démêler l'écheveau du fil informationnel contenu dans l'intégralité du système d'information de l'organisation pour laquelle il travaille. Les éditeurs de logiciels ont mis au point des systèmes permettant d'indexer et de hiérarchiser l'information pour la rendre accessible au moment voulu. Appelés km (knowledge management) ou ged (gestion électronique des documents), ces systèmes se sont récemment développés pour apporter une logique et un sens aux requêtes des utilisateurs ; il s'agit là d'une évolution stratégique qui pourrait remettre en question certaines entreprises de veille.

Mais au préalable, il a été indispensable de qualifier les formes d'informations à disposition avant d'espérer les organiser. Celles-ci ont été classées en deux parties essentielles : les informations structurées et non structurées.

I. L'information structurée

Les documents structurés sont au format électronique et conçus selon une structure logique qui permet de les hiérarchiser dans des bases de données. La nécessité de stocker des informations en nombre toujours croissant a vu naître ce qu'on appelle les bases de données, puis les erp*, et les grands noms de ce monde sont Oracle et sap. Ce sont des progiciels qui permettent de gérer les processus d'une entreprise en intégrant l'ensemble de ses fonctions, dont la gestion des ressources humaines, la gestion comptable et financière, l'aide à la décision, mais aussi la vente, la distribution, l'approvisionnement et le commerce électronique. Une information est considérée comme structurée si elle peut être traitée automatiquement par un ordinateur et non nécessairement par un humain. Elle est donc stockée dans un sgbd* et possède une architecture interne cohérente et prédéfinie.

II. L'information non structurée

À l'opposé, l'information non structurée ne possède aucune structure fixe prédéfinie et son contenu est variable. Les données ne sont ni classées ni identifiées comme celles que l'on trouve dans les courriels, dans les ordinateurs (documents Office ou pdf), les flux rss, les pages Web, les blogs, forums et sur des supports audio ou vidéo. Elle représente aujourd'hui 80 % de l'information d'une organisation. Les données essentielles concernant un client ou un projet en cours de réalisation seront très logiquement transmises par e-mail, de même que les échanges entre les différents acteurs internes et externes de l'entreprise. Il est donc essentiel pour chaque intervenant d'avoir connaissance d'une part de ces informations et d'autre part, d'identifier l'ensemble des collaborateurs qui participent directement ou non

à l'élaboration du projet. Il s'agit donc d'informations non partagées, situées sur les postes de travail que les organisations doivent impérativement traiter.

III. Les applications de type Desktop Search

Parmi les solutions proposées par les éditeurs, on trouve des applications dites de Desktop Search : elles sont souvent gratuites et permettent aux utilisateurs de bénéficier d'un moteur de recherche interne. Selon le cabinet Markess International, elles représentent 47 % des solutions utilisées.

1. Google Desktop Search

Cette application de bureau permet d'effectuer des recherches en texte intégral sur les e-mails, les fichiers, la musique, les photos, les chats, la messagerie électronique Gmail, les pages Web consultées et bien plus encore. En rendant votre ordinateur accessible aux recherches, Google Desktop permet d'exploiter les informations et évite d'avoir à organiser manuellement fichiers, e-mails et favoris.

Après son installation, Google Desktop commence à indexer le contenu de l'ordinateur. Cette indexation initiale n'est effectuée qu'à partir du moment où l'ordinateur est inactif pendant plus de trente secondes, afin de ne pas ralentir la machine, mais elle peut demander plusieurs heures (en fonction du nombre d'éléments à indexer). Google Desktop actualise l'index en ajoutant les nouveaux messages à mesure de leur réception, les fichiers à chaque modification et les pages Web consultées.

2. Copernic Desktop Search

Copernic Desktop Search (cds) permet de trouver instantanément des fichiers, courriels et fichiers joints, peu importe où ils se trouvent sur le disque dur du système. L'application permet de retracer des fichiers de type Word, Excel et Powerpoint de Microsoft, pdf d'Adobe et multimédia notamment. cds est offert en trois versions : la version Home destinée aux particuliers, la version Professional destinée aux professionnels, et la version Corporate destinée aux entreprises.

3. Windows Desktop Search

Windows Search est un outil conçu pour vous aider à retrouver facilement les documents, mp3, e-mails, rendez-vous et autres documents disséminés sur un pc. Il se présente sous la forme d'un champ de recherche placé dans votre barre des tâches et permet ainsi de rechercher le document souhaité. La recherche peut être filtrée en spécifiant un emplacement ou un type de fichier particulier.

4. Exalead Desktop Search

Exalead Desktop Search permet aux utilisateurs de retrouver instantanément tout document présent sur les disques durs internes ou externes, qu'il s'agisse d'un document Office, d'un message, d'une pièce jointe, d'un contact, d'un rendez-vous, etc. L'outil permet aussi à chacun de chercher en explorant sa propre arborescence de fichiers, qu'elle soit sur la messagerie ou sur le disque dur.

IV. Comparatif

Encore faut-il savoir exactement ce que l'on cherche, toute la difficulté est là. Il ne suffit pas de connaître les mots-clés, les termes associés sont aussi importants que les dates ou le format du document souhaité.

Un test sur les différents outils proposés par les éditeurs démontre l'importance de ces nuances qui font toute la qualité d'un moteur de recherche conçu et adapté à l'esprit de l'utilisateur.

En effet, les solutions proposées par Microsoft, Google ou Copernic livrent un résultat sur le mot-clé saisi ; or ce mot-clé est la plupart du temps insuffisant et nous sommes souvent à court d'idée pour compléter la requête. Par exemple, dans le cadre de la réalisation d'un projet sur un nouveau modèle de voiture avec le client Dupont, il sera peut-être nécessaire de rechercher une information traitée avec lui. Cependant, si M. Dupont est client de l'entreprise depuis cinq ans, si le projet en cours est le troisième, il risque d'y avoir beaucoup de données le concernant. Peut-être aussi l'information recherchée concerne-t-elle d'autres collaborateurs, qu'elle soit sous un format spécifique (pdf, Excel, Word, e-mail, etc.), mais a été diffusée à la suite d'une réunion qui a eu lieu l'année précédente.

La requête à saisir serait : *Dupont*.

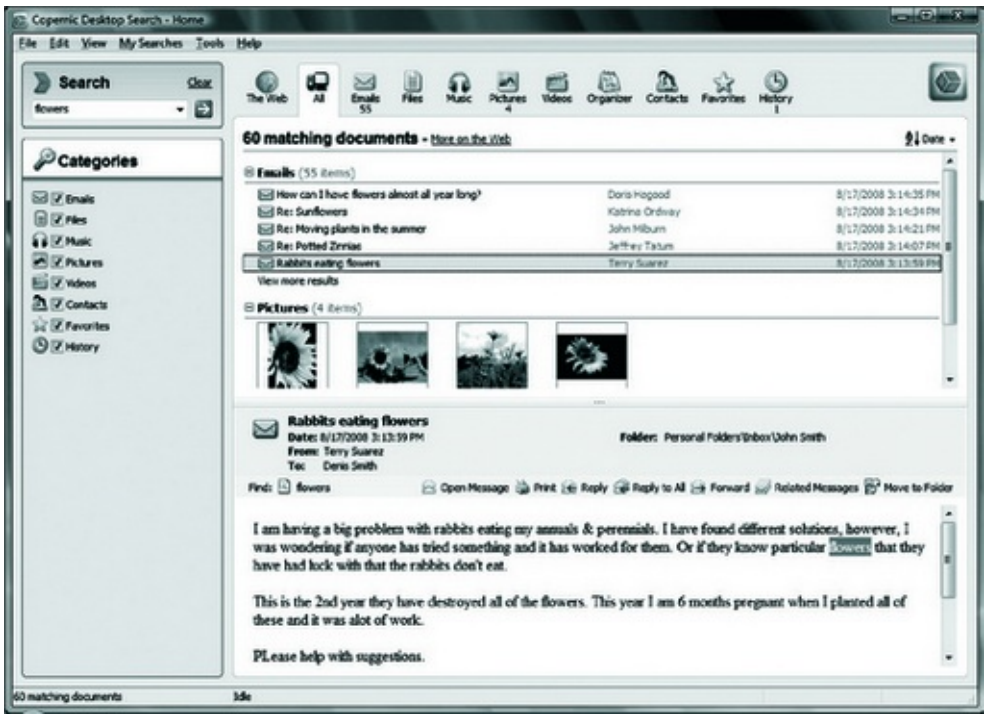
Suivie, éventuellement, d'une limitation dans le temps pour la période visée (Google).

Les résultats afficheraient l'ensemble des données dans lesquelles le mot *Dupont* apparaîtrait. Ce qui obligerait l'utilisateur à consulter l'intégralité des informations avant de retrouver celle qui l'intéresse. Alors que les moteurs de recherche disposent les résultats en vrac ou par type de fichier – Copernic, par exemple, indique que le mot-clé est présent sur un certain nombre d'e-mails, de fichiers Word ou autres –, Exalead met à disposition un système de navigation qui assiste l'utilisateur à l'aide de suggestions diverses.

Concrètement cela se présente sous la forme d'une interface dans laquelle sont proposées toutes sortes de critères pour faciliter la recherche. Ainsi la requête initiale *Dupont* aboutit à une liste cliquable qui définira :

- les sources (e-mails, disques durs, bureau, favoris, etc.) ;
- l'arborescence des répertoires et le nombre de documents recensés ;
- les types de fichiers (Acrobat, message, Word, Jpeg, etc.) ;
- les auteurs ;
- les destinataires (pour les e-mails) ;
- les termes associés ;
- la date (triée par année, puis, lorsqu'elle est sélectionnée, par mois, et enfin par jour) ;
- la taille des fichiers ;
- les langues (anglais, allemand, espagnol, etc.) ;
- le multimédia (vidéo, images, etc.).

Chaque clic modifie la page de résultats en conséquence ; en outre, d'autres critères d'affinage sont proposés. L'utilisateur a alors l'opportunité de scanner les résultats pour juger d'un simple coup d'œil de la pertinence de ce qui est affiché grâce à un résumé automatique, à la présence de vignette, de la prévisualisation avancée et du surlignage qui permettent, sans avoir à ouvrir le document, de décider rapidement de son intérêt.



V. Les risques liés aux outils de recherche Desktop

Ces solutions de recherche sur les ordinateurs comportent des risques importants en termes de sécurité. Si Exalead Desktop Search s'installe sur le poste et n'en bouge plus, certaines autres applications n'en font pas de même et les failles ont des conséquences qui inquiètent les responsables en sécurité des systèmes d'information.

Il existe des solutions globales qui couvrent aussi bien l'information interne qu'externe à l'organisation et ce, quelle que soit leur localisation (poste de travail, bases de données, applications métier, sites Internet, serveur de messagerie, etc.). Ou d'autres qui permettent à l'utilisateur d'héberger ses données sur des serveurs tiers pour y accéder depuis n'importe quel endroit dans le monde. Si ces outils sont parfois indispensables pour l'organisation, ils peuvent représenter une faille de sécurité particulièrement préoccupante pour la confidentialité des données traitées.

L'eff (Electronic Frontier Foundation) a, par exemple, alerté les utilisateurs et leur a fortement suggéré de ne pas utiliser l'application Google Desktop qui représente un risque significatif d'exploitation des données personnelles par le renseignement américain. Cette application, comme celle d'Apple (MobileMe), permet de stocker et d'indexer l'intégralité des informations (e-mails personnels et professionnels, carnet d'adresses, contrats, informations financières, messageries instantanées, historique de navigation sur Internet) des utilisateurs sur leurs propres serveurs.

Une fonctionnalité (optionnelle) retient particulièrement l'attention : la possibilité de rechercher des contenus sur plusieurs ordinateurs simultanément. Elle est censée faciliter la vie des personnes ayant l'habitude de travailler sur plusieurs postes. Lorsque cette option est activée, les copies texte des

documents et l'historique Internet d'un ordinateur sont automatiquement transférés vers le poste sur lequel l'utilisateur travaille. Et lorsqu'il tape une requête, les informations contenues sur les deux ordinateurs peuvent ainsi être passées au crible. La société garantit que toutes les copies des documents présents sur ses serveurs disparaissent au bout de trente jours, une durée largement suffisante pour récupérer des informations sensibles selon l'eff. Par ailleurs, si un pirate arrive à obtenir le login* et le mot de passe d'un utilisateur, il pourra avoir accès à l'ensemble des données – ce qui est plus efficace et discret que de pénétrer le système d'information d'une organisation. En cas d'incident, les données risquent d'être rendues publiques ou accessibles à des personnes non autorisées. En mars 2009, des usagers de Google Docs, la suite bureautique de Google, ont reçu un courriel les informant que leurs documents pourraient avoir été partagés par accident. Selon la firme américaine, un bug a causé le partage involontaire de documents avec des correspondants non autorisés mais avec lesquels il y aurait déjà eu des échanges. Pour limiter les dégâts, tous les usagers affectés ont reçu une liste des documents en cause et tous les fichiers potentiellement touchés par accident ont perdu leurs droits d'accès [1]. Cet incident, dont l'évaluation des dommages n'a atteint que 0,05 % des abonnés selon Google, laisse entrevoir l'impact que pourrait avoir une interruption de service due à une panne ou à un attentat. Les utilisateurs dépendants de ces solutions seraient les victimes directes d'une fermeture de leur compte ou de leur accès aux partages de fichiers avec leurs collaborateurs, la grande majorité n'ayant certainement pas établi de plan de secours.

En fait, de nombreux outils (comme Autonomy avec Richard Perle assistant du secrétariat à la Défense sous Reagan et Chairman du Defense Policy Board Advisory Committee sous l'administration Bush) ont des liens avec des services de renseignements ou de défense, ce qui, en soi, est parfaitement logique mais peut se révéler délicat pour les utilisateurs étrangers détenteurs d'informations sensibles.

VI. La recherche d'informations dans une organisation

Les organisations font face à des enjeux au cœur desquels la gestion de l'information devient stratégique : multiplication des sources d'informations (internes et externes), structurées (bases de données) et non structurées (e-mails). Ce foisonnement de données exige la mise en place de solutions spécifiques permettant de mieux accéder, rechercher, analyser et diffuser l'information afin d'en faciliter le partage et l'exploitation en entreprise.

Dès lors se pose un problème fondamental : comment retrouver l'information, ou, plus précisément, comment exploiter désormais ces contenus en information ? Car ce foisonnement des informations, des données, des serveurs et des bases pour les stocker pose aujourd'hui plusieurs problèmes complexes :

- les bases de données sont, comme leur nom l'indique, des entrepôts de données ; celles-ci sont conçues et optimisées pour ranger les données et non pour accéder aux données. Dès lors, pour interroger une base, la question doit s'insérer temporairement dans le processus naturel d'entreposage de la base (sa fonction première) de stockage et de déstockage des données ; elle doit parfois patienter avant de pouvoir s'insérer. Interroger une base de données par les processus classiques perturbe donc son fonctionnement, le processus est intrusif, et donc cher ;

- de surcroît, les bases, destinées à ranger des quantités impressionnantes de données, sont dessinées selon des critères simples : nom, numéro de téléphone, numéro de facture, produit. Les interroger « classiquement » n'a de sens qu'en fonction des quelques questions simples qui ont présidé à la définition de leur structure.

Mais les questions que les entreprises sont amenées à se poser n'ont aucune raison de correspondre à la structure même des bases. Par exemple, la question : « Que sais-je de mon client ? » n'est pas adaptée à la logique des bases de données, or cette question est clé. De surcroît, une question de ce type ne ressort pas à l'interrogation d'une seule base : les grandes institutions multiplient les bases en fonction des problématiques qui s'imposent à elles. Comme les données sont éparpillées (éparses voire de plus en plus externes, c'est-à-dire présentes sur le Web, dans un blog, un forum, etc.), la « représentation » de la question, simple à l'origine : « Que sais-je de mon client ? », pour telle grande institution, nécessite 700 requêtes, 700 questions différentes à plusieurs bases de données, ce qui constitue une réponse démesurée en nombre.

Comme, de surcroît, chaque question est intrusive, le problème devient progressivement insoluble, le coût de la réponse monte en flèche et l'on arrive alors aux limites d'une technologie.

Des éditeurs ont mis au point des solutions logicielles dont les fonctionnalités cumulent celles de la ged et du km avec divers niveaux de sophistication selon les objectifs désirés par les entreprises – celles-ci cherchant parfois à fournir un outil performant destiné aux veilleurs pour alimenter les employés directement sur les sujets qui les concernent, à moins qu'elles ne préfèrent donner à l'ensemble des salariés les moyens d'effectuer des recherches directement sur l'ensemble des informations dont elles disposent tout en tenant compte des droits d'accès mis en place par le rssi* ou le dsi*.

Le point le plus important est donc le sens donné à une requête, c'est la raison pour laquelle les moteurs de recherche internes ne sont pas *forcément intéressants* pour l'utilisateur final, ils ne savent pas encore donner de sens au résultat. (Remarque : on cherche toujours un document mais on a vraiment besoin d'une information, l'avenir des moteurs de recherche d'entreprise, c'est de fournir de l'information prête à l'usage, qui supporte une décision, de l'information dite productive.)

Dans une organisation, le ranking de type Google peut donc difficilement s'appliquer, le collaborateur a besoin d'une réponse *cohérente avec son profil, sa position, son métier, une réponse* qui associe les résultats en une forme logique et exploitable dans un contexte précis.

Notes

[1] Source : Techcrunch.

Chapitre VIII

Un savoir-faire français

La plupart des métiers liés au traitement de l'information électronique sont complémentaires et hiérarchisés dans leurs activités. Les éditeurs français sont particulièrement performants dans le domaine du traitement de l'information, les solutions proposées sont à la pointe de la technologie.

I. Numérisation

Jouve : *Digitalisation industrielle de documents et de données, traitement d'image, diffusion électronique*. Bien des organisations possèdent encore des informations archivées sur un support papier. Jouve est spécialisé dans la dématérialisation d'information qui permet de numériser l'ensemble des données d'une entreprise pour qu'elles puissent être ensuite indexées et hiérarchisées. Les activités de Jouve sont réparties sur plusieurs sites de production en France (Paris, Mayenne, Orléans, Rennes, Lens et Nancy) et à l'étranger : en Europe, aux États-Unis, au Sénégal et en Chine. L'entreprise réalise 40 % de son chiffre d'affaires à l'export.

II. Recherche sémantique

Pertimm : *Moteur de recherche*. Pertimm est un éditeur de logiciel implanté en France et aux États-Unis (Nevada et Californie), spécialiste des moteurs de recherche multimédia. Créée en 1997 par trois ingénieurs experts en intelligence artificielle, l'entreprise propose notamment des études personnalisées sur les systèmes d'information des clients concernant principalement l'activité de recherche d'informations.

1. Recherche « intelligente » d'informations

Les moteurs de recherche Web accessibles aujourd'hui au grand public fonctionnent à partir de l'analyse de mots-clés ou de caractères, et se fondent sur une approche quantitative des données – ces dernières étant classées selon leur popularité (le page rank) et leur date de création. Malgré un fort taux d'utilisation, les internautes reconnaissent que ces outils restent limités : obligation d'exprimer avec exactitude et en peu de mots le sujet de la recherche, présence de silences (pages pertinentes pour une requête donnée non remontées par le moteur, voire non indexées), manque de pertinence des réponses…

Dans le cadre du projet Infom@gic, Pertimm a élaboré un moteur de recherche Web 3.0 nouvelle génération qui s'appuie sur des fonctionnalités inédites. L'éditeur français entend ainsi démontrer l'intérêt et la valeur ajoutée d'un moteur linguistique et sémantique en matière de recherche d'informations.

Destiné au plus grand nombre, ce nouvel outil proposera une alternative efficace aux grands acteurs du marché de la recherche Web, particulièrement grâce à sa technologie des « Pertimmiseurs » déjà mise en

place avec succès depuis plus de cinq ans sur les Intranets de ses clients. À la différence des moteurs existants, l'outil Pertimm intègre un traitement sémantique par Pertimmiseurs. La prise en compte de la linguistique, des ontologies, des entités nommées et des concepts, ainsi qu'une navigation dans les résultats par co-occurrences lui permettent d'offrir aux utilisateurs une palette de réponses beaucoup plus large et pertinente que les résultats obtenus habituellement.

Il offre ainsi à l'utilisateur la possibilité d'effectuer une recherche sur un corpus de mots étendu. Actuellement, au-delà de trois ou quatre mots, celle-ci se solde par un échec ou par des réponses non pertinentes. Avec Pertimm, plus l'utilisateur entre de mots, meilleurs sont les résultats obtenus.

Autre nouveauté : la prise en compte de l'aspect qualitatif des sources plutôt que le page rank. Cela permet aux internautes d'avoir accès à des articles fondamentaux en dépit d'une date de création très ancienne ou d'une consultation réduite.

Ce nouveau moteur de recherche est donc particulièrement adapté pour des recherches spécifiques ou lorsque l'utilisateur ne sait pas comment formuler ce qu'il recherche.

2. Exalead

Moteur de recherche. Éditeur français présent en Europe (Royaume-Uni, Benelux, Allemagne, Espagne, et Italie), ainsi qu'aux États-Unis, Exalead est le seul acteur du secteur à disposer d'une technologie qui sert à la fois son moteur Web grand public (huit milliards de pages, un million de visiteurs uniques et des fonctionnalités inédites qui ont inspiré les plus grands acteurs du Web, comme la recherche image ou la navigation par termes associés) et sa suite logicielle Exalead CloudView destinée aux entreprises. Ce savoir-faire hybride fait d'Exalead un éditeur considéré comme très innovant sur ses propositions d'expérience utilisateurs (ergonomie, interactivité et multimédia) et particulièrement performant sur la notion de passage à l'échelle, un prérequis indispensable à la création de services d'accès à l'information de demain (volume de données en croissance exponentielle, hétérogénéité des sources et des formats dans l'entreprise et aussi sur la Toile, importance du temps réel, hausse du trafic, etc.).

Avec plus de 200 clients dans le monde, tout secteur confondu – du public aux services –, Exalead essaie aujourd'hui de repenser l'usage des moteurs de recherche, non plus uniquement sous la forme du : « Je pose une question, j'ai des documents », mais sous la forme : « Je pose une question, je dispose d'une information contextualisée et prends une décision fondée » et exploite de manière étonnante sa technologie d'index pour concevoir des applications métiers sur des corpus aussi bien structurés que non structurés, dans la logistique, les centres d'appels ou encore les services online.

Exalead fait partie du projet Quaero, innove sur la Toile avec son labs, et poursuit son expansion, fort de ses réussites commerciales, de ses implémentations réussies et de la reconnaissance des grands analystes américains du secteur.

La technologie d'Exalead repose sur la combinaison à l'échelle industrielle de modules de traitement linguistique, à la fois statistiques et sémantiques, et sur une plate-forme très puissante et suffisamment flexible pour intégrer des technologies tierces (data mining, reconnaissances visuelle et vocale) et ainsi s'adapter à tous les usages dans l'entreprise et sur le Web.

III. Text mining

Le Text mining représente l'automatisation du traitement d'une masse importante de données textuelles non structurées, l'objectif étant d'extraire les principales tendances. À partir de là, peuvent être répertoriés, de manière statistique, les différents sujets évoqués, afin d'adopter des stratégies plus pertinentes, résoudre des problèmes et saisir des opportunités commerciales.

Le Text mining a différentes fonctions : adopté comme filtre de communication, il aide à la classification des mails, notamment des spams ; comme optimiseur de recherche d'informations, il améliore la consultation des documents et facilite la gestion des connaissances en proposant des résumés de texte par la sélection, etc.

1. Temis

Traitement de l'information. Créé en septembre 2000, Temis est présent en France et à travers ses filiales, en Allemagne, en Italie, au Royaume-Uni et aux États-Unis. L'entreprise développe des solutions logicielles de Text mining qui permettent d'optimiser le traitement de l'information du texte libre en données analysables pour l'extraction d'informations ou le classement automatique de documents. Temis est la première société à avoir adapté ses solutions à la fois aux problématiques spécifiques des entreprises (découverte scientifique, enrichissement de contenu, suivi de réputation et voix du client, et intelligence économique) et aux besoins de plusieurs secteurs (sciences de la vie, industrie, édition et médias, Sécurité nationale).

2. Arisem

Analyse sémantique et Text mining. Arisem est un éditeur de logiciels spécialisé dans le traitement automatisé du langage. Filiale du groupe Thales, l'entreprise conçoit, développe et commercialise des solutions de Text mining et d'analyse automatisée du langage, centrée sur l'analyse et l'organisation de l'information. Elle permet une analyse fine du contenu textuel par l'exploitation du sens des phrases et non pas uniquement par une recherche de mots-clés.

IV. Veille

Une veille efficace permet d'accéder, au moment opportun, à une information adaptée aux besoins et aux attentes. Elle doit favoriser l'amélioration des pratiques de travail et contribuer à la professionnalisation individuelle et collective. Organisée et structurée en plusieurs étapes, elle concourt à améliorer le temps de recherche et à livrer des renseignements utiles (informations traitées : ordonnées, synthétisées et d'appropriation facile). La veille permet surtout d'analyser les bruits du Net, les tendances, les réputations des organisations privées ou institutionnelles et de détecter les menaces comme les opportunités.

1. Digimind

Veille stratégique. Digimind est une plate-forme de veille stratégique d'entreprise : elle permet à ses utilisateurs de disposer des informations les plus à jour sur leur domaine d'activité afin de prendre les décisions les plus éclairées possible. Elle propose des solutions destinées à ceux qui font et organisent la veille (documentalistes, responsables de la veille) et à ceux qui utilisent ses résultats (comités de direction, responsables marketing, commerciaux, chercheurs, etc.). Elle permet également de déployer et

d'animer des équipes et projets de veille stratégique, particulièrement sur les médias du Web social comme Twitter, Facebook ou Del.icio.us.

Ainsi, il est possible d'interroger ou de surveiller de nombreux bouquets de médias sociaux :

- sites de microbloguing : Twitter, Jaiku… ;
- sites de réseaux sociaux : Facebook, Linked In… ;
- sites de partages multimédias : Flickr, DailyMotion, YouTube… ;
- sites d'avis consommateurs : Ciao, ToLuna… ;
- sites de partages de documents : Del.icio.us, SlideShare…

Sans oublier les forums Web et groupes de discussions qui hébergent chaque jour des centaines de milliers de conversations, essentielles pour une veille stratégique complète.

2. Ami Software

Recherche d'information et valorisation des connaissances. ami Software est un éditeur de logiciels spécialisé dans les produits d'acquisition, de gestion et de traitement d'informations textuelles destinés aux organisations. Ses solutions sont au centre de projets de veille, d'intelligence économique, de gestion et de valorisation de connaissances et d'analyse d'opinions. Go Albert, qui a pour marque commerciale ami Software, a été créé par un groupe d'industriels du monde de l'aviation et de l'électronique. Cette origine a conduit la société à adopter des processus de conception et de fabrication issus de ces industries.

Le moteur de recherche Discovery d'Ami Software permet de retrouver des informations textuelles dispersées dans des courriels, des documents ou des bases de données sans en connaître la localisation exacte. Cette recherche s'effectue soit en langage libre à partir d'une question ou d'un texte, soit en langage booléen. L'objectif affiché de ce moteur de recherche est de limiter le « bruit » – les réponses qui ne présentent pas ou peu d'intérêt – et le « silence », c'est-à-dire les réponses oubliées. La présentation des résultats peut être effectuée par sources ou sous forme de listes fusionnées fédérant les différentes sources. À noter que ce moteur de recherche tolère certaines fautes d'orthographe grâce aux recherches par similarité, et offre la reconnaissance d'acronymes.

Go Albert a notamment développé un logiciel offrant une panoplie complète d'outils qui permettent d'écouter, d'organiser, d'analyser les opinions exprimées par les internautes sur le Web 2.0 qui fournissent les synthèses indispensables aux décideurs.

3. kb Crawl

Solutions de veille. Fondé en 1995, kb Crawl est spécialiste de l'identification, de la collecte et du traitement de l'information, notamment dans les domaines à haute valeur ajoutée de l'informatique financière et de l'informatique industrielle. L'éditeur propose des solutions de veille allant de la collecte d'informations à la diffusion en passant par le filtrage, le traitement et la capitalisation des données. Son objectif est de mettre en place un point d'observation unique pour surveiller tout type d'information sur le

Web visible (sites, blogs, forums, etc.) et invisible (bases de données, moteurs de recherche, réseaux sociaux, etc.).

V. Linguistique : l’analyse sémantique verticale

Lingway commercialise plusieurs solutions permettant l’implémentation de moteurs de recherche spécialisés ou verticaux (essentiellement dans le domaine médical et les ressources humaines). Par exemple Lingway Medical Suite (lms) est une solution d’analyse de textes médicaux et de recherche dans les bases de données médicales s’appuyant sur un dictionnaire électronique et des composants sémantiques adaptés au langage médical. Lingway Medical Dictionary Encoder (lmde) est un moteur de recherche français et anglais dédié à la pharmacovigilance. Il permet un codage automatique dans la nomenclature internationale Meddra (Medical Dictionary for Drug Regulatory Activities) des phrases (verbatims) décrivant les effets secondaires des médicaments, les antécédents et les pathologies associés.

Lingway a développé d’autres outils d’aide au codage notamment pour la classification internationale des maladies (cim) et la classification commune des actes médicaux (ccam), destinée au codage des actes médicaux.

Chapitre IX

Perspectives

I. En entreprise

Le volume croissant de données provenant d'Internet (forums, blogs, sites Web…) draine une quantité considérable d'informations non exploitables, ce qui entraîne une utilisation limitée de ces données par les entreprises. Pourtant, ces informations diverses et variées sont indispensables pour une bonne connaissance des prospects et des clients. Si elles ne sont pas prises en compte, la vision des entreprises ne peut être que partielle. Dans le cas contraire, les perspectives sont extraordinairement riches. Prenons un exemple avec les laboratoires pharmaceutiques qui s'interrogent sur le moyen de travailler sur les effets secondaires alors qu'ils ne disposent pas assez rapidement du recul nécessaire pour obtenir ces informations ou qu'ils ne peuvent identifier par des tests cliniques. On peut alors imaginer qu'il y a potentiellement, dans la totalité des forums dans le monde, de l'information qui traite de ces effets secondaires. Là est l'intérêt : définir à quel instant ce qui, dans la masse de données de tous les contenus disponibles, prend un sens pour l'entreprise. Nous sommes dès lors bien loin de la simple recherche de document. C'est d'ailleurs là que se trouve l'avenir du moteur de recherche : non plus dans la recherche menant à une liste de données, mais dans la reconnaissance d'une suite d'événements qui conduiront à un résultat.

En d'autres termes, à quel moment la réconciliation de tera-octets de données – maîtrisées ou non, issues d'organisations, de partenaires, de consommateurs, de gouvernements – par des moteurs de recherche sera utilisée pour fournir du sens au résultat, au sens de l'entreprise. Par exemple, une firme pourra lier intelligemment les données internes et externes concernant un produit, mais aussi les informations issues du marketing, de la finance, des ventes, ainsi que les retours des clients (appréciations, échanges de mails, retours marchandise, réassorts, etc.), les commentaires des consommateurs (forums, blogs, réseaux sociaux, presse, etc.), l'image de l'entreprise sur la Toile, de ses cadres, de ses concurrents, le positif et le négatif : en fait, tout ce qui permet d'avoir une vision globale sur ce qui fonctionne ou non dans l'entreprise et pour quelles raisons.

II. Dans le monde du Search

Google indexe à ce jour plus de 20 milliards de pages Web, un nombre qui croît régulièrement, au point que nous serons bientôt saturés par ces données inextricables. Désormais, les moteurs de recherche s'efforcent de proposer de nouvelles fonctionnalités qui permettent aux internautes d'accéder aux informations dont ils ont besoin, avec des méthodes de recherche différentes, plus intuitives et plus rapides. L'enjeu est de taille : hormis le fait qu'il est impossible dorénavant de combattre Google sur un pied d'égalité, les moteurs tentent par tous les moyens de grignoter des parts de marché en augmentant le trafic qui représente l'unique argument de vente pour les annonceurs. La recherche s'oriente donc vers de nouveaux services à valeur ajoutée et des méthodes innovantes comme la recherche thématique.

Ouvert en mai 2009, Bing, le moteur de recherche de Microsoft lui a permis de récupérer 2 % de parts de marché deux semaines après son lancement soit de 9,1 à 11,1 % aux États-Unis. Le successeur de Live Search bouscule la mise en page traditionnelle des moteurs avec pour fond une photo en couleur qui change quotidiennement. Les informations sont classées par sous-rubriques, plutôt que sous forme de listes de liens Web. Si l'internaute tape le mot Afrique, il peut choisir entre plusieurs thèmes, tels que vacances, tourisme, immobilier, culture, etc. Par ailleurs, des outils d'aide à la décision seront intégrés pour réserver, par exemple, un vol au meilleur prix ou choisir un restaurant.

L'analyse des requêtes quotidiennes de millions d'internautes permet d'anticiper leurs futures demandes et de générer des termes associés pour faciliter le choix d'une page de résultats. L'affichage de ces mots-clés classés par importance est situé dans une colonne grisée à gauche de l'écran.

Microsoft mise sur les outils d'aide à la décision et annonce dans un communiqué : « Face à la croissance exponentielle de l'information disponible sur Internet, l'expérience de recherche peut être aujourd'hui considérablement améliorée. Seule une requête sur quatre apporte une réponse satisfaisante du premier coup et 15 % des recherches sont abandonnées faute d'avoir trouvé une réponse. »

La recherche des images offre aussi une nouvelle interface avec un affichage agrandi lors du survol de la souris. Les paramètres de recherche affinée sont sur la colonne de gauche avec la possibilité de sélectionner, entre autres, la taille, la disposition et le style de la photo. Si l'internaute le désire, il peut consulter l'image dans son contexte avec une prévisualisation du site dans laquelle elle est publiée, puis retourner aux résultats d'un simple clic sur un lien situé au-dessus du frame. Dans ce cas, ce sont les images qui se projettent dans la colonne de gauche et qui, à leur tour, peuvent être sélectionnées pour une prévisualisation.

Les acteurs cherchent à se démarquer et à séduire les internautes avec des tests en ligne des nouveaux concepts au fur et à mesure de leur réalisation. Ces applications originales connaissent un fort développement et répondent aux besoins des consommateurs lassés des listes de liens interminables.

Miiget permet de visualiser graphiquement l'ensemble des individus reliés à une personnalité. Une expérience qui démontre la capacité à créer de l'information structurée autour de personnes extraites directement à partir du Web.

Google Squared propose une liste sous forme de tableau dont les colonnes et les lignes peuvent être modifiées au gré de l'internaute. Générées automatiquement, ces listes permettent d'organiser l'information, de la personnaliser et de la sauvegarder.

Voxalead indexe les vidéos en s'appuyant sur une technologie Speech to text, qui découpe les vidéos et permet d'accéder directement aux passages citant les mots recherchés. Une nouvelle version plus riche, plus verticale, nommée Voxalead News propose, elle, de rechercher directement dans les news vidéos et podcasts des plus grandes chaînes d'information (cnn, France 24, abc, bbc, Europe 1, etc.) [1].

L’ultime moteur de recherche reste donc à inventer. Il saurait lier l’information issue de sources diverses, interne et externe, structurée ou non, mais surtout organiser le contenu pour lui donner un sens. Idéalement, il serait apte à fournir de l’information généraliste comme de l’information verticale sur un sujet précis. Capable de retenir toutes les préférences de l’utilisateur, il ne serait pas pour autant inquisiteur ni espion.

Véritable outil d’aide à la décision, il offrirait, sous forme graphique, l’ensemble des moyens technologiques disponibles sur une seule plate-forme. Mais cette solution présente un problème considérable de confidentialité et de sécurité. Comment garantir l’intégrité des données personnelles sur un support qui connaît tout de son utilisateur ? Pris entre la qualité des fonctionnalités et la préservation des informations privées, l’internaute doit encore travailler sur plusieurs supports, au risque de perdre un peu de temps.

Notes

[1] Voxalead News repose sur Exalead CloudView et intègre le module de transcription Speech to text élaboré par Vecsys, dans le cadre de leur collaboration sur le projet Quaero.

Conclusion – Dépendance : le risque d'une rupture

Nous l'avons vu, l'apprentissage de la gestion de l'information à l'heure d'Internet est devenu un enjeu essentiel, et ce d'autant que le monde dépend aujourd'hui totalement des technologies de la communication. Au risque de perdre la maîtrise des informations que l'on transmet s'en ajoute ainsi un autre : aucune organisation ne peut durablement se retrouver coupée de connexion à Internet ni retourner à l'époque du fax et du courrier postal.

Nous sommes dans l'instantanéité et ne pouvons plus attendre plusieurs jours pour recevoir un document par la poste. Les organisations commencent enfin à prendre conscience du risque d'une rupture des réseaux due à une interruption volontaire ou accidentelle des systèmes d'information. L'Estonie – l'un des pays les plus connectés d'Europe, dont la majorité des services ne sont accessibles qu'en ligne (notamment les services bancaires) – a réalisé les conséquences du tout-numérique sans avoir préalablement mesuré l'impact d'une attaque sur l'infrastructure informatique. Début 2007, le gouvernement estonien décide de déplacer une statue érigée à la gloire des soldats de l'armée rouge en plein centre de Tallinn vers un cimetière qui lui semble plus approprié. Le 27 avril, jour du déboulonnage, les sites Internet et les serveurs estoniens (gouvernement, Bourse, banques, assurances) ont été la cible d'attaques Ddos* massives isolant le pays le plus connecté d'Europe pendant quarante-huit heures. L'intervention d'experts internationaux et de l'otan a été nécessaire pour lutter contre des assauts numériques d'une intensité encore jamais vue.

Les effets dévastateurs sur les systèmes d'information dus à l'arrêt des services en ligne (temps d'intervention pour les réparations, pertes financières, etc.) ont engendré une soudaine panique au plus haut niveau des États, de l'otan et de l'Union européenne. Depuis cette date historique [1], chaque conflit ou simple désaccord entre deux pays génère systématiquement des attaques numériques sur les sites gouvernementaux mais aussi sur les extensions (par exemple le «.fr » pour la France) afin de faire le plus de victimes possible [2].

En dehors des cyberconflits ou des guérillas numériques, les risques de rupture de connexion s'appliquent aussi au secteur des entreprises, fournisseurs d'accès ou de services avec lesquels la dépendance atteint un niveau préoccupant.

Notes

[1] Bien que les attaques interétatiques remontent au début des années 2000, le conflit Estonie-Russie est devenu la référence en matière de cyberconflit.
[2] Laurence Ifrah , « Analyse de la première attaque massive d'un État », *La Revue de la Défense nationale*, novembre 2007.

Glossaire

La connaissance de quelques termes est indispensable à la compréhension du Web 2.0 :

api

L'Application Programming Interface est une interface qui permet à deux programmes informatiques de communiquer entre eux grâce à des standards communs. Par exemple, la Google api est un kit de développement logiciel disponible librement qui permet de créer de nouvelles applications utilisant directement la base de données des pages indexées par Google, par le biais d'un service Web.

Crowdsourcing

Le crowdsourcing pourrait se traduire par « l'approvisionnement par la foule » ; cela consiste à mutualiser les ressources de milliers d'internautes en utilisant leur temps disponible pour créer du contenu, résoudre des problèmes ou autres et proposer ainsi des services à moindre coût. Des sites comme istockphoto proposent plusieurs millions de photographies d'amateurs à des prix variant de 1 à 20 dollars maximum. ExpertExchange offre les services d'un public amateur ou professionnel en informatique. Les internautes présentent un problème technique et la communauté essaye de trouver la solution, le gagnant remporte des points qu'il pourra utiliser pour « monnayer » les services d'un spécialiste lorsqu'à son tour il aura besoin de conseils.

Ddos

L'attaque Ddos permet à celui qui la lance de saturer par de fausses requêtes les serveurs de la cible au point de les rendre indisponibles pendant une durée variant entre quelques heures et plusieurs jours. Les serveurs configurés pour recevoir un certain nombre de demandes ne sont plus en mesure de transmettre les données requises si le volume de connexions devient excessif. Il est très difficile voire impossible, dans certains cas, de contrer ce type d'attaque d'où son succès.

dsi

Directeur des systèmes d'information.

erp

L’Enterprise Resource Planning est un progiciel qui permet de gérer l’ensemble des processus de l’entreprise comme la comptabilité, les ressources humaines, les ventes, etc.

Folksonomie

Cette fonctionnalité phare du Web 2.0 est un système de classification collaborative décentralisée spontanée, selon des mots-clés attribués par les utilisateurs. Les sites Del.icio.us et Flickr utilisent ce système pour classer leur contenu. Ce qui permet d’améliorer la recherche et de fournir aux autres le contenu de sa propre collection de ressources.

Log

Journal d’événements. Les logs recensent toutes les activités ayant eu lieu sur les systèmes d’information ou sur un site. Par exemple pour les connexions à un site, les logs vont en conserver la date, l’heure et l’origine. Ces journaux permettent d’identifier la source d’un dysfonctionnement mais peuvent également être utilisés à des fins marketing ou de renseignement.

Login

Identifiant de connexion qui peut être un pseudo, l’e-mail ou le nom de l’utilisateur et qui est généralement suivi d’un mot de passe.

Mashups

Les mashups (mixage en français) permettent de construire des services en ligne basés sur l’assemblage de données provenant de plusieurs sources. Ces services peuvent à leur tour être enrichis de contenus générés par les utilisateurs. Le mashup permet également au producteur d’un contenu de le proposer à la communauté des développeurs qui pourront ensuite le formater et le valoriser sous une autre forme. C’est une forme d’externalisation de l’intelligence collective des internautes et des développeurs.

Phishing

Le phishing consiste à faire la copie graphique d’un site légitime, principalement un site financier ou de vente en ligne, pour obtenir les identifiants de l’utilisateur et lui pirater son compte bancaire.

Podcasting

Terme issu de la combinaison de iPod et de broadcasting qui consiste à mettre en ligne sur un site, un fichier audio (mp3 ou mp4) ou vidéo (videocasting) au format numérique auquel peut s'abonner un internaute pour le télécharger sur un baladeur et l'écouter à sa convenance.

Post

Terme employé couramment pour indiquer qu'un billet a été posté (publié) sur Internet.

Proxy

Un proxy est un serveur informatique qui a pour fonction de relayer des requêtes entre un poste client et un serveur Web. Ce dernier identifiera la connexion du proxy et permettra à l'internaute de rester anonyme dans une certaine mesure.

rss

Voir Syndication.

rssi

Responsable de la sécurité des systèmes d'informa-tion.

sgbd

Le système de gestion de base de données est un ensemble de programmes qui permettent l'accès à une base de données.

Social-bookmaking

Le bookmarking social consiste à partager ses favoris avec les internautes qui présentent les mêmes centres d'intérêts. L'autre avantage est que l'ensemble des favoris est conservé sur une plate-forme en ligne qui permet la consultation depuis n'importe quel ordinateur ou navigateur.

Syndication rss et Atom

La syndication Web permet d'extraire d'un site Web ou d'un blog du contenu régulièrement mis à jour. Elle peut être libre et gratuite, ou payante pour les sites utilisant les contenus mis à disposition. Le plus souvent, ils sont distribués par le biais de fils d'information. Ce procédé a été démocratisé grâce aux blogs dont le succès repose en partie sur l'usage de fils rss (Really Simple Syndication) ou Atom, deux formats populaires de flux d'information. Les sites abonnés à un fil donné peuvent enrichir leurs propres contenus : en contrepartie, ils renvoient les internautes vers le site émetteur du flux.

Tags

Les tags sont des étiquettes réalisées par la communauté des internautes. Ils apparaissent dans un tag cloud (ou nuage de mots) dont la taille est proportionnelle à la fréquence de leur utilisation et permet de visualiser les qualificatifs les plus employés sur un sujet donné.

Taxonomie

À l'origine, il s'agit de la science de la classification des êtres vivants : elle a pour objet de les décrire et de les regrouper en entités appelées taxons (familles, genres, espèces, etc.), afin de pouvoir les nommer et les classer. C'est aussi la science des lois et règles qui déterminent l'établissement des méthodes et systèmes de classement (systématique).

Wiki

Un wiki est un site Internet dynamique qui permet aux internautes de rédiger et de modifier librement des articles qui interagissent entre eux sur un site Web. La philosophie du wiki veut que les informations soient modifiables par tous, mais il est possible de restreindre la visualisation ou l'édition des pages par un mot de passe. La modération se fait *a posteriori*, grâce, notamment, à la fonction « derniers changements » qui permet à tout moment de revenir à une version antérieure. Le mot wiki vient de l'hawaiien. Il signifie « vite ». Le premier wiki a été créé en 1995 par Ward Cunningham.

Bibliographie

Bibliographie et Webographie

- Andrieu O. , *Trouver l'info sur le Web*, Paris, Eyrolles, 2001.
- Calishain Tara et Dornfest Rael , *Google à 200 %*, O'Reilly, 2005.
- cehd *Des réseaux et des hommes*, Paris, L'Harmattan, 2000.
- Dornfest R. , Bausch P. et Calishain T. , *Google Hacks*, O'Reilly, 2006.
- Faligot Roger et Kauffer Rémi , *Histoire mondiale du renseignement*, Paris, Robert Laffont, 1993.
- Huyghe François-Bernard , *Maîtres du faire croire*, Paris, Vuibert, 2008.
- Long J. , *Google Hacking*, Paris, Dunod, 2005.
- Malbreil Xavier , *La Face cachée du Net*, Omniscience, 2008.
- Sacks Risa , *Super Searchers go to the Source*, Reva Bash, 2001.
- Sadin Éric , *Surveillance globale*, Climats, 2009.
- www.abondance.com
- www.amazon.com
- www.amisw.com
- www.aol.com
- www.aolstalker.com
- www.arisem.com
- www.copernic.com
- www.cybion.fr
- www.dailytech.com
- www.del.icio.us
- www.derspiegel.de
- www.digimind.com
- www.directioninformatique.com
- www.economist.com
- www.eff.org
- www.exalead.com
- www.google.com
- www.informaticien.be
- www.journaldunet.com
- www.jouve.fr
- www.lefigaro.fr
- www.lemonde.fr
- www.lingway.com
- www.liveplasma.com
- www.paperblog.com
- www.rankspirit.com
- www.reuters.com
- www.temis.com

- www.theguardian.com
- www.thekbuzz.com
- www.thewired.com
- www.thomascrampton.com
- www.touscomplices.com
- www.twitter.com
- www.uhb.fr
- www.wmaker.net
- www.wordle.net
- www.yahoo.com
- www.zdnet.com