Le syndrome de Big Brother : qu'est-ce que les moteurs connaissent de nous ?

Retour au sommaire de la lettre

Dans la rubrique "Big Brother", les données personnelles détenues par les moteurs de recherche sont au cœur de toutes les préoccupations depuis l'injonction du ministère de la Justice américain d'accéder aux données détenues par certains moteurs sur les requêtes effectuées par leurs utilisateurs sur un laps de temps donné. Mais quelles sont donc les informations que les outils de recherche sont susceptibles de détenir sur nous ? Le risque d'intrusion dans nos vies privées est-il si important ? Nous vous offrons ici un panorama des principales données qui sont recueillies par ces outils.

Vous voulez savoir ce que pense quelqu'un ? Quels sont ses goûts favoris ? Ses habitudes de consommation ? Regardez donc les requêtes qu'il a effectuées sur son moteur de recherche préféré, cela vous en dira certainement long sur ses centres d'intérêt… Est-ce possible techniquement ? Une telle idée s'apparente-t-elle à de la paranoïa ? Il est très difficile de le savoir… Dernièrement, la justice américaine a en tout cas demandé à consulter une liste des requêtes effectuées sur plusieurs moteurs de recherche nationaux lançant, par là même, une polémique sur l'éventuelle utilisation des données stockées par les moteurs à des fins de cybersurveillance. Mais de nombreux outils proposés par Google notamment (Gmail, Desktop Search, Personalized Search entre autres) avaient déjà suscité le débat...

Si certains outils de recherche ont mis à la disposition de l'Etat américain une partie des informations demandées (en les rendant anonymes), Google fait quant à lui de la résistance depuis plusieurs mois pour protéger la vie privée de ses utilisateurs et ses secrets commerciaux.

Google serait donc plus susceptible de protéger la vie privée de ses utilisateurs que Yahoo ou MSN ? Pas si sûr selon le Times britannique qui souligne qu'à l'instar d'autres moteurs américains, Google fait fréquemment l'objet de critiques de la part des défenseurs des libertés qui l'accusent notamment de coopérer avec le gouvernement chinois pour appliquer une censure stricte des sites dans ce pays (http://www.timesonline.co.uk/article/0,,11069-2002169,00.html).

Mais pourquoi tant d'attention sur le cas de Google alors même que, de l'aveu de la plupart des analystes, tous les moteurs majeurs (y compris Yahoo, MSN ou A9) conservent autant de données privées que Google sur leurs utilisateurs ? *"Google est sur le point de voler la vedette à Microsoft en ce qui concerne son potentiel de violation de la vie privée de ses utilisateurs. Les consommateurs ont beaucoup de mal à accepter cette situation car elle est en rupture totale avec l'image forte de confiance de ce moteur"*, explique Chris Hoofnagle du "Electronic Privacy Information Center" (voir http://news.com.com/Googles+balancing+act/2100-1032_3-5787483.html).

Rappel historique : en août 2005, les autorités américaines ont demandé à consulter une liste des requêtes des utilisateurs

Le 25 août 2005, Google, MSN et Yahoo ont reçu une injonction du ministère de la justice américain de leur fournir certaines des données qu'ils collectent sur leurs utilisateurs, et plus particulièrement les requêtes effectuées par les internautes sur une période de deux mois.

Objectif affiché du DoJ : étudier ces données en vue d'évaluer l'importance des contenus illégaux et de relancer une loi de 1998 destinée à protéger les enfants de la pornographie sur le Web (Child Online Protection Act). Cette loi a été suspendue il y a deux ans par la cour suprême qui l'avait alors jugée inconstitutionnelle.

Selon le New York Times, MSN, AOL et Yahoo ont accepté de fournir des données au DoJ, tandis qu'Ask Jeeves n'a pas été sollicité. Google s'y refuse quant à lui avec force, pour plusieurs raisons : le respect de la vie privée de ses utilisateurs, la protection de son patrimoine technologique et des secrets commerciaux, ainsi que la charge que ferait peser cette requête gouvernementale sur ses serveurs. Autre explication avancée par ce moteur : certaines requêtes pourraient à elles seules révéler l'identité de ceux qui les ont formulées (en imaginant qu'ils aient, par exemple, effectué une recherche sur leur numéro de carte de crédit).

Côté DoJ, on affirme que les informations fournies ne peuvent en aucun cas permettre d'identifier les internautes et que le respect de leur vie privée est donc assuré.

Les défenseurs des libertés civiles ne l'entendent pas de cette oreille. A l'instar de Beth Givens, directrice d'une association de protection des libertés civiles de San Diego, ils sont nombreux à estimer que *"Les moteurs de recherche sont un objectif très tentant pour le gouvernement"* (voir http://www.liberation.fr/page.php?Article=352805). Selon Ray Everett-Church (un consultant spécialisé sur les questions de protection de la vie privé), *"C'est exactement ce que les défenseurs de la vie privé redoutent depuis longtemps. Le pire que l'on puisse craindre est que ces énormes bases de données soient désormais ouvertes à n'importe qui avec un papier officiel. Si ils perdent cette bataille, de nombreux consommateurs y regarderont maintenant à deux fois avant de laisser Google entrer dans leur vie"* (voir http://www.siliconvalley.com/mld/siliconvalley/13657386.htm).

Non sans humour, les auteurs d'un article publié sur le weblog américain BoingBoing (auquel collabore le spécialiste des outils de recherche John Battelle), demandaient également dernièrement à AOL, Yahoo et MSN de rendre publiques les données qu'ils ont fourni pour prouver qu'elles ne sont pas privées (http://www.boingboing.net/2006/01/30/boingboing_search_pr.html) !

Mais, au fait, les requêtes des utilisateurs d'un moteur sont-elles réellement d'ordre privé ? Elles le sont très certainement si l'on considère que *"la plupart des gens qui effectuent des requêtes sur des moteurs ont l'impression d'avoir une sorte de conversation confidentielle avec le moteur qu'ils utilisent. Ils n'imaginent en aucun cas que leur requête puisse ensuite faire le tour du monde"*, estime Danny Sullivan (SearchEngineWatch.com). Ce qui effraie certains observateurs, c'est aussi bien entendu le fait que l'historique des recherches d'un individu sur un moteur peut offrir une définition assez précise de qui il est (ses sujets de prédilection, croyances religieuses, préférences sexuelles et peut-être même ses problèmes médicaux lorsqu'il a effectué des recherches à ce sujet).

Quelles sont les données personnelles recueillies par les moteurs de recherche "classiques" ?

Il convient ici d'effectuer une distinction entre les moteurs de recherche "classiques" et les grands portails de recherche – comme Google Yahoo et MSN – qui proposent des outils de communication en complément de leurs services de recherche d'information.

Tous les moteurs majeurs recueillent quatre types d'informations sur leurs utilisateurs :

- Leur **adresse IP** (seul votre fournisseur d'accès Internet a la possibilité de vous identifier avec cette adresse),
- Les **requêtes qu'ils ont effectuées**,
- Les **cookies** (qui indiquent si votre navigateur s'est déjà connecté auparavant),
- Les **comptes utilisateurs** lorsque les utilisateurs en ont un (dans ce cas, le moteur dispose peut-être aussi de leur adresse e-mail).

Selon François Bourdoncle, dg d'Exalead, les données recueillies *"sont très élémentaires et non nominatives"* puisque son moteur ne dispose *"pas de mécanisme qui permettrait d'identifier les utilisateurs"* (comme, par exemple, un compte webmail). Il ajoute que ces données *"sont utilisées à des fins statistiques pour évaluer le type de trafic, son origine géographique, la pertinence du moteur, etc"*. Elles sont donc difficilement exploitables pour autre chose que des statistiques d'utilisation.

Dans sa "politique de protection des données personnelles", le moteur de recherche Mirago précise lui qu'il "*ne contrôle pas, ne garde pas en mémoire, n'essaie pas d'identifier les noms, adresses e-mail, adresses professionnelles ou personnelles des personnes effectuant des recherches sur ses sites. A cet égard, toute recherche est faite dans le plus grand anonymat*" (http://fr.mirago.com/apropos/respetinfoperso.htm). Le moteur ajoute qu'il collecte certaines *"informations ne permettant pas de vous identifier quand vous visitez nos sites directement ou par le biais de la barre de recherche Mirago. Ces informations sont stockées dans un fichier log et incluent votre adresse IP, le type de navigateur Internet utilisé, la date et l'heure de la visite, la date et l'heure de votre recherche ainsi qu'un ou plusieurs cookies"*. Fabrice Mégange, son responsable France, explique que ces données sont exclusivement exploitées pour le système de geotargetting des annonces publicitaires.

Idem chez Seekport, qui indique sur son site que *"certaines informations non nominatives sont automatiquement collectées et conservées : le User Agent (le nom de votre navigateur), votre adresse IP, la date et l'heure d'accès à notre site, le http code (l'état de connexion à notre serveur), le cas échéant, le site web par lequel vous êtes arrivé sur Seekport"* (http://www.seekport.fr/help/privacy.html). Le moteur précise qu'il n'enregistre *"que des informations anonymes, dans le but d'établir des statistiques"*.

Trois questions à Laurent Baleydier (métamoteur KartOO)

Quelles sont les principales données qu'un outil comme le votre peut recueillir sur ses utilisateurs ?
Sur KartOO, nous stockons :
- les 100 dernières recherches de l'utilisateur avec leurs dates,
- les 100 dernières recherches ayant abouties (c'est à dire pour lesquelles un site a été cliqué),
- les 100 derniers sites cliqués ou personnalisés (avec la date de dernière visite, une couleur, une note et un commentaire),
- une dizaine de carte sauvegardées ou crées par l'utilisateur.

Combien de temps Kartoo conserve-t-il les données qu'il recueille sur les usages des internautes, si il recueille de telles données ?
Ces données sont anonymes et stockées sur le poste de l'utilisateur jusqu'à ce qu'il les efface. Il les gère une à une grâce au "Kapitalyser" qui se trouve dans le menu historique à gauche de la case de recherche. Elles ne passent JAMAIS par le serveur de KartOO, nous n'avons pas de base de données ce qui garantit le respect de la vie privée.

A quoi servent ces données lorsqu'elles sont exploitées ?
A personnaliser les résultats. Par exemple, en fonction du mot tapé, KartOO peut suggérer une requête qui a déjà abouti (dans les dossiers à gauche de la carte).
Il peut aussi mettre en évidence les sites qui ont été cliqués ou personnalisés avec des pictogrammes différents sur la carte.

Notons ici que certains analystes pensent toutefois que l'augmentation des capacités de stockage et la baisse des coûts risque d'inciter tous les administrateurs de sites Web et de moteurs de recherche à accroître la quantité de données qu'ils sauvegardent sur leurs visiteurs.

Quelles sont les données détenues par les grands portails de recherche et de communication ?

Google place systématiquement sur les PC de ses utilisateurs (qui ne bloquent pas ce type de fichier) un cookie qui répertorie à la fois leurs requêtes, la date à laquelle elles ont été effectuées et les résultats qu'ils ont sélectionnés. Ce cookie n'est pas nominatif mais il identifie les caractéristiques de leur système et leur adresse IP, permettant donc leur identification si besoin. Google Watch (http://www.google-watch.org/), un site d'activistes américains, prétend lui que la durée de vie du cookie de Google est inconnue et que les informations sont conservées indéfiniment.

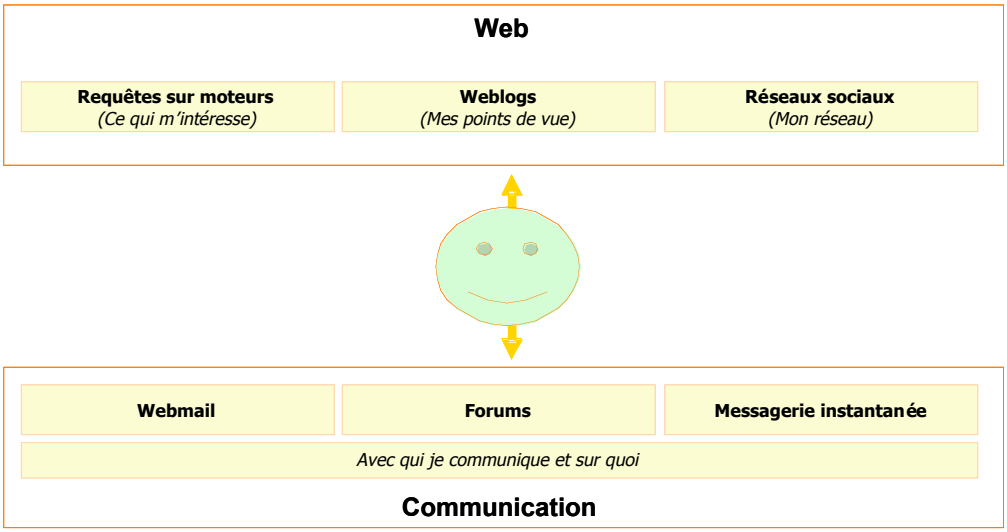
Google enregistre également automatiquement dans des fichiers logs sur ses serveurs *"les pages dont l'affichage est demandé lors de la consultation du site par les internautes"*.

Ces données sont ensuite utilisées pour contrôler et améliorer l'efficacité du moteur. Google se réserve le droit de combiner ces données *"avec les informations issues d'autres services Google ou de tiers, dans le but d'en rendre l'utilisation plus agréable et, le cas échéant, de proposer des contenus personnalisés"* et de *"partager des informations non personnelles sous forme collective avec des tiers"* (http://www.google.fr/intl/fr/privacy.html).

Les inquiétudes quant à la protection de la vie privée ont surtout tendance à se focaliser sur la combinaison des informations obtenues par les services de recherche et celles qui proviennent des services de communication.

Si l'on prend le cas de Google, certains estiment que les recoupements des informations obtenues grâce à ses différents systèmes de recherche, de mail (Gmail), de messagerie instantanée (Google Talk), de shopping (Froogle), de recherche sur PC (Google Desktop) ou encore son réseau social (Orkut) pourraient à terme créer l'une des bases de données les plus détaillées du monde sur les goûts et les habitudes de communication des individus. Si on ajoute à cela des outils comme le projet Gdrive (disque dur personnel déporté sur les serveurs de Google), on se rend compte que les risques d'intrusion sont réels s'ils ne sont pas "bornés" à un moment ou à un autre...

Type de données détenues par les grands portails de recherche et de communication



Comme en témoigne le tableau ci-dessous, Google n'est pas le moteur qui collecte le plus d'informations personnelles sur ses utilisateurs lors de leurs inscriptions. Pour l'ouverture d'un service Gmail, Google demande un login et une adresse mail alors que Yahoo et MSN demandent, eux, davantage d'informations comme le nom, la date de naissance et le code postal de l'utilisateur.

Comparaison des données utilisateurs pouvant être recueillies par les trois principaux moteurs (liste non-exhaustive)

	Yahoo	MSN	Google
Recherche Web			
Adresse IP	x	x	x
Terme recherché	x	x	x
Date de la recherche	x	x	x
Cookies	x	x	x
Services proposés	x	x	x
Recherche sur PC	x	x	x
Recherche images	x	x	x
Recherche vidéo	x	x	x
Recherche audio	x		
Recherche locale	x	x	x
Recherche d'actualités	x	x	x
Recherche shopping	x	x	x
Recherche finance	x	x	
Recherche voyages	x	x	

	Yahoo	MSN	Google
Recherche musique	x	x	x
Groupes de discussion	x	x	x
Réseaux sociaux	x	x	x
Weblogs	x	x	x
Compte utilisateur			
Nom	x	x	
Prénom	x	x	
Sexe	x	x	
Adresse e-mail	x	x	x
Industrie	x		
Poste occupé	x	x	
Date de naissance	x	x	
Code postal	x		
Pays	x	x	x
Webmails			
e-mails	x	x	x
Contacts	x	x	x
Messagerie instantanée			
Messages	x	x	x
Contacts	x	x	x

Perspectives

Les informations d'ordre privé recueillies par les moteurs de recherche sont variées et relativement inexploitables par des services tiers à l'heure actuelle, en tout cas pour ce qui est des mots clés saisis dans les formulaires de recherche. N'oublions pas, non plus, que les utilisateurs peuvent "éliminer" régulièrement les cookies sur leurs PC… Même si le "problème" des fichiers logs reste entier.

La demande des autorités américaines d'accéder aux requêtes des utilisateurs a certainement le mérite d'ouvrir un large débat sur la possible nécessité de contrôler les conditions de collecte, de détention et d'exploitation de ces données "privées" par les moteurs.

En effet, si aucun moteur ne vend actuellement les données qu'il collecte sur ses utilisateurs et si la fourniture d'informations à des institutions gouvernementales semble bien encadrée dans la plupart des pays, n'est-il pas nécessaire de garantir aux utilisateurs le maintien à venir de leur anonymat pour qu'ils continuent d'utiliser ces outils en toute confiance ? Dans ce cas, la problématique est à la fois centrée sur les mots clés saisis et sur les informations détenues dans la cadre du compte utilisateur nécessaire pour accéder à certains et outils. Ces données privées représentent un véritable trésor, s'il est possible de les croiser avec les historiques de recherche des utilisateurs. Quelles garanties les moteurs nous donneront-ils à courte et moyenne échéance qu'elles ne seront pas utilisées à grande échelle pour des activités mercantiles ou autres ? L'avenir des moteurs de recherche actuels n'est-il pas lié à la réponse à cette question ?