Peut-on tout confier à Google ?

Article publié le 14 Novembre 2008
Source : LE MONDE.FR Stéphane Foucart
Taille de l'article : 3130 mots

En dix ans d'existence, Google a tellement grandi qu'il a fini par se rendre incontournable. Notre courrier, notre mémoire, bientôt notre dossier médical… chaque jour, les serveurs de l'entreprise accumulent de nouveaux détails sur notre intimité. Mais comment le géant Google gère-t-il nos données personnelles ?

Depuis plusieurs semaines, il soupçonnait si fortement sa femme d'infidélité qu'il ne parvenait plus à obtenir la moindre érection. Il s'était mis en quête d'un conseiller matrimonial, tout en cherchant le meilleur moyen de confondre son épouse. Sonoriser la voiture ? Installer des caméras de surveillance pour bébés dans l'appartement ? Pister l'activité de son téléphone portable ? Le jour où il avait découvert qu'elle le trompait bel et bien, et que l'amant était une amante, il avait sombré dans l'alcool. Au fond de sa déprime, il imaginait mettre un contrat sur les deux femmes. Ce résident de Floride avait cherché un contact auprès d'une mafia au Portugal, son pays d'origine. Désespéré, il avait fini par s'enquérir d'une aide au suicide, avant de se raviser, et de décider de quitter le continent américain.

Comment connaissons-nous autant de détails privés de cette triste histoire vraie ? Quelle autre trahison, en plus de celle de son épouse infidèle, a valu à cet homme de voir ainsi publiquement exposée son intimité ? Celle de son analyste ? D'un membre de sa famille ? D'un proche confident ? Non : de son moteur de recherche. Car ce qui fait le canevas de ce drame domestique n'est rien d'autre, ni rien de plus, que la suite de mots-clés qu'il a recherchés, jour après jour et pendant trois mois, sur le Web.

En août 2006, AOL a divulgué accidentellement un fichier contenant les logs de 658 000 utilisateurs, collectés entre les mois de mars et de mai de la même année. Les logs, c'est-à-dire l'historique des mots-clés recherchés sur Google depuis le portail d'AOL, accolés à la date et à l'heure de leur recherche. Chaque internaute n'est certes identifié que par un numéro – celui de l'époux trahi est 14162375 – mais bien vite, des journalistes et des blogueurs se sont amusés à rechercher, parfois avec succès, des identités réelles derrière les suites de chiffres…

Aux Etats-Unis, l'affaire a frappé les esprits. Car ce n'est pas seulement "une base de données d'intentions humaines", selon l'expression du journaliste américain John Battelle (La Révolution Google, éd. Eyrolles, 282 p., 19,90 €), que révèlent ces logs : c'est aussi et surtout un immense catalogue d'angoisses, de pulsions, de terribles secrets, de noirs fantasmes ou de perversions cachées… L'utilisateur 11574916 cherche "cocaïne dans l'urine". Le numéro 1515830 se demande "comment dire à votre famille que vous êtes victime d'inceste". Un autre, le numéro 59920, veut savoir "à quoi ressemble un cou après avoir été étranglé" et quelle "corde utiliser pour cravater quelqu'un"…

Simplicité et harmonie

Autant le dire : Google est parfaitement innocent. La société de Mountain View (Californie), qui ne jouait d'autre rôle que celui de prestataire de service pour le portail d'AOL, n'a jamais eu la moindre intention de dévoiler ces logs. Elle fait même plutôt figure, dans cette affaire, de victime collatérale. Car la faute d'AOL a focalisé une énorme attention sur les moteurs de recherche et, bien sûr, sur le premier d'entre eux. Mais que représentent les quelque 20 millions de logs rendus publics par le fournisseur d'accès américain, face aux milliers de milliards stockés en permanence dans les serveurs de Google, cette entreprise à laquelle nous confions chaque jour davantage de nos goûts, de notre mémoire, de notre intimité ?

Ces histoires de logs, Peter Fleisher en est un peu fatigué. Agacé, presque. Chez Google, le responsable mondial de la protection des données a choisi de travailler à Paris. Tout un symbole : la France est le premier pays à s'être doté, en 1978, d'une loi sur l'informatique et les libertés. Cet Américain avenant au français irréprochable reçoit dans les locaux de la filiale française du géant américain, au deuxième étage d'un immeuble haussmannien, face à l'Opéra-Garnier. Design simple et épuré, le blanc domine, tout juste teinté, par-ci par-là, des quatre couleurs d'un logo que plus grand-monde n'ignore sur la planète. Des bureaux à l'exacte image de la page d'accueil historique de Google : simplicité et harmonie.

"Nous sommes les meilleurs élèves de la classe", tranche Peter Fleisher. De toute évidence, sur la gestion de ces fameux logs, il a raison. Même si d'aucuns jugent cela insuffisant. En avril 2008, le groupe Article 29 (ou G29), qui rassemble toutes les autorités européennes de protection des données personnelles, a publié une analyse juridique des activités de recherche sur Internet. Le G29 recommande en conclusion que les logs générés par les recherches en ligne ne soient pas conservés plus de six mois. "Nos durées de conservation étaient de dix-huit mois et nous venons de les réduire à neuf mois, rétorque M. Fleisher. Et c'est une durée de conservation bien inférieure à celles pratiquées par nos concurrents", principalement Yahoo ! et Microsoft.

Même si le G29 reconnaît cet effort, des points de friction demeurent. "Dans la réponse qui nous a été fournie et qui a été rendue publique, Google récuse le fait que ses traitements de données personnelles [ses fichiers, en somme] soient soumis à la législation européenne, explique ainsi Gwendal Le Grand, expert à la direction des affaires juridiques de la Commission nationale de l'informatique et des libertés (CNIL). Il considère que le responsable du traitement des informations est Google Inc. – basé au Etats-Unis – et que le rôle des filiales nationales se limite à la vente de publicité." C'est pourtant bien la publicité qui est au centre de tous les soupçons. "Google ne récupère pas de données personnelles pour profiler des utilisateurs et leur proposer de la publicité en fonction de leurs caractéristiques, assure Peter Fleisher. C'est une légende urbaine. Nous ne pratiquons que la publicité contextuelle." (lire l'encadré ci-contre).

Savoir ce que sait Google : la question est d'autant plus cruciale que la société n'est plus seulement, depuis longtemps, un moteur de recherche. Google est devenu tout autre chose – chose qui reste d'ailleurs à définir. L'entreprise délivre et stocke du courrier électronique (Gmail), propose des services de cartographie (Google Earth) et de calcul d'itinéraires (Google Maps), ausculte et décortique le trafic des sites Internet (Google Analytics), compile la presse (Google News), gère des agendas en ligne (Google Calendar), édite des blogs (Blogger), indexe la littérature savante (Google Scholar), stocke des albums photo (Picasa) et des vidéos (YouTube), scanne et conserve des livres (Google Book Search), gère des carnets de santé électroniques (Google Health)… Google chasse aussi depuis peu sur les terres de Microsoft en proposant un traitement de

texte et un tableur en ligne (Google Documents) qui permet de s'affranchir de la sacro-sainte et fort coûteuse suite logicielle Microsoft Office. Google sait tout, Google s'occupe de tout. Depuis quelques semaines, il est même embarqué sur un smartphone (G1) concurrent de l'iPhone d'Apple et fait pièce à Internet Explorer et Firefox avec son nouveau navigateur (Chrome)… Autre versant de son succès : Eric Schmidt, son PDG, est désormais un des principaux conseillers du nouveau président Barack Obama.

Liste non exhaustive.

Au siège de Google France, on rappelle volontiers que nombre de sociétés – banques, assurances, opérateurs de télécommunications – disposent elles aussi de données touchant à l'intimité de leurs clients. Certes. Mais jamais autant de données privées n'ont été assemblées sur une seule et même plate-forme, sur un système informatique intégré.

Où sont localisées ces informations ? Où sont stockés ces logs, mais aussi toutes ces précieuses bribes de vie quotidienne, ces requêtes, ces correspondances, ces documents ? Où repose cette "base de données d'intentions" ? "Il est difficile de répondre à cette question, dit Peter Fleisher. Je comprendrais que les utilisateurs trouvent confortable de se dire : Tiens, mes données sont à tel endroit . Mais la vérité est que celles-ci ne sont pas stockées sur un serveur situé sur un site particulier : elles le sont sur un réseau d'un grand nombre de serveurs situés sur un grand nombre de sites, eux-mêmes dispersés dans plusieurs pays. Tout est d'ailleurs stocké en deux endroits au minimum pour qu'en cas de coupure d'électricité sur un site, on ne perde pas de données…"

Ce que décrit Peter Fleisher a un nom : le cloud computing, cette "informatique nébuleuse" dans laquelle toutes les données des utilisateurs (courriers, documents, albums photo, vidéos, etc.) ne sont pas nécessairement inscrites sur le disque dur de leur ordinateur mais sont stockées en ligne, sur des réseaux de gros serveurs. La tendance naturelle étant d'externaliser aussi les logiciels, pour ramener l'ordinateur personnel à une boîte presque vide, simple terminal d'accès à une sorte de cerveau global, ainsi que l'explique Nicholas Carr dans son dernier ouvrage (The Big Switch, éd. Norton). Comme on branche son grille-pain au réseau électrique sans se soucier de savoir comment l'électricité est produite, on connecterait un ordinateur ou un smartphone au Google cloud pour accéder à ses données et à ses logiciels sans se soucier de savoir où et comment ils sont conservés.

Où ? Comment ? On touche là au grand secret de Google. Son infrastructure informatique est un réseau de plusieurs dizaines de centres de données, chacun alignant de milliers de serveurs. Le tout forme à la fois la mémoire et l'intelligence de cette machine universelle qu'est devenue Google au fil du temps, la plus puissante plate-forme informatique de stockage et de calcul distribué jamais conçue. Dire de Google qu'il est le Web ne relève pas d'un abus de langage ou même d'une description imagée de la réalité : ses ordinateurs parcourent inlassablement le Web pour le copier en intégralité. En même temps qu'elle aspire l'ensemble de la Toile pour pouvoir la passer à la moulinette de ses algorithmes, la "machine Google" répond en une fraction de seconde à des milliards de requêtes journalières, chacune ciblant quelques mots éparpillés dans une dizaine de milliards de pages…

Capacité de résistance

Interroger Google sur l'infrastructure matérielle qui permet cette performance expose à une fin de non-recevoir. De combien de serveurs la société dispose-t-elle ? Dispersés dans combien de centres de données ? Ces chiffres sont parmi les mieux gardés de l'industrie informatique. Une bonne source, qui tire ses estimations de discussions avec

des ingénieurs de la firme de Mountain View, avance le nombre de 900 000 serveurs répartis dans 20 à 50 centres de données – dont au moins trois en Europe : en Irlande, en Belgique, aux Pays-Bas (The Economist du 23 octobre avance le chiffre de 2 millions de serveurs). Les localisations de certains de ces centres sont publiques. Les visites, elles, sont formellement proscrites. Pour raisons de sécurité, bien sûr. Mais aussi, confie Peter Fleisher, parce que “voir la manière dont est constitué l’intérieur d’un centre de données fournirait déjà des informations à la concurrence”.

Faire confiance à une société qui tient secret les lieux dans lesquels elle entend garder toute l’information du monde ? “Du point de vue de la confidentialité des données, la localisation d’un centre de données n’est pas très importante”, répond Peter Fleisher. Ce qui compte c’est, en effet, la curiosité des gouvernements, des forces de l’ordre et des services de renseignements. Et surtout la capacité de Google à y résister. Car, potentiellement, toutes les données stockées en ligne peuvent être, un jour où l’autre, opposées à leur “propriétaire”. Ce n’est pas qu’une question de principe. Shi Tao, un journaliste chinois condamné en avril 2005 et qui purge une peine de dix ans d’emprisonnement, en sait quelque chose. Titulaire d’un compte de courrier électronique sur Yahoo !, il avait vu sa correspondance privée divulguée auprès de ses accusateurs par la société américaine. Google, dont l’ambition est d’indexer et de conserver toute l’information du monde – domaines public et privé confondus –, pourrait-il se permettre le risque d’un pareil écart ? A l’évidence non. “Je suis fier que les blogueurs birmans aient décidé de publier leurs photos des événements récents [les manifestations des moines bouddhistes] sur la plate-forme d’édition de Google [Blogger], martèle Peter Fleisher. Ils savaient qu’ils ne risquaient rien.”

Pour autant, indique l’expert et consultant américain Stephen Arnold, l’un des meilleurs connaisseurs de Google, la société de Mountain View “reçoit beaucoup de requêtes des forces de l’ordre américaines et étrangères”. “Mais la majorité d’entre elles est simplement ignorée, ajoute Stephen Arnold. Il n’y a pas de système mis en place par Google pour répondre à ce genre de demandes.” D’un point de vue pratique et juridique, dit en substance M. Arnold, l’espionnage est donc compliqué. Mais il l’est aussi d’un point de vue technique, puisque “le système n’a pas été pensé ni conçu pour répondre à ce genre de demandes”.

Peu de cas sont connus. Mais, par exemple, les forces de l’ordre brésiliennes ont récemment tenté de se renseigner, via une demande à la firme californienne, sur les clients d’un réseau de vente en ligne d’objets nazis – et Google Inc. leur a opposé une fin de non-recevoir, la propagande nazie n’étant pas, au pays du free speech, considérée comme illicite. Il demeure, en somme, beaucoup plus compliqué à un service de police de récupérer quoi que ce soit de la forteresse Google que de faire placer un quidam sur écoutes ou, plus simplement, de requérir de son fournisseur d’accès à Internet (FAI) la liste des sites sur lesquels il s’est connecté – puisque la législation française impose à tous les FAI la conservation de ces informations.

Même quand il ne s’agit pas de requêtes émanant de services de police, Google se montre souvent très sourcilleux. A la suite d’un litige pour violation présumée massive du droit d’auteur, la société Viacom a récemment requis de Google qu’il fournisse les logs de consultation de certaines vidéos mises en ligne par sa filiale YouTube. Ce qui fut fait, mais après “anonymation” de ces logs… Ces précautions ne procèdent pas seulement d’un souci moral. Mais aussi de la crainte de perdre un leadership qui, comme on le rappelle chez Google France, “repose entièrement sur la confiance des utilisateurs”. Et ce d’autant plus que Google ambitionne depuis peu de fournir ses services de gestion de l’information

aux entreprises qui, de plus en plus sensibilisées aux questions d'intelligence économique, n'ont aucune envie de voir leurs petits et grands secrets sur la place publique.

De futures crispations sécuritaires pourraient changer cette donne. Mais, pour l'heure, Google prend un soin particulier à préserver les données des utilisateurs de toute curiosité indue. Fin de l'histoire ? Pas tout à fait. Car si préserver les données des curieux est encore possible, éviter leur perte par accident pourrait se révéler, dans l'avenir, un peu plus délicat. "Google n'est pas infaillible", prévient Stephen Arnold, qui rappelle que voilà seulement quelques semaines, un "plantage" en règle a rendu le service Gmail indisponible pour de nombreux utilisateurs, pendant près de trente heures. Gérer un système de près de 900 000 machines interconnectées soumises à des dizaines de millions de requêtes quasi simultanées n'est pas chose aisée.

D'autant que le Google cloud n'est pas un dispositif statique. Il se reconfigure en permanence, épouse les pulsations d'Internet. En fonction de l'alternance entre le jour et la nuit, par exemple, c'est-à-dire en s'adaptant à l'activité régionale des internautes, les données valsent et sont transférées d'un bout à l'autre de la planète. Pour alléger la charge ici, l'alourdir là où cela ne portera pas à conséquence… Le vrai " péché " d'indiscrétion de Google se situe d'ailleurs là. "Google se moque de la nature des données personnelles proprement dites, explique Stephen Arnold. Ce qui est tracé, ce sont les méta-données générées par les activités des internautes." Ce qui intéresse Google, ajoute-t-il, "c'est, par exemple, de savoir si les taux de consultation des horaires d'avion sont plus élevés à proximité des aéroports pour pouvoir ainsi s'adapter".

Autre raison de se montrer indiscret : Google doit s'assurer, en permanence, que les clics sur ses liens publicitaires sont de "véritables" clics. C'est-à-dire qu'ils ne relèvent pas de clics machinalement exécutés pour produire artificiellement un revenu indu… Cette chasse à la "fraude au clic" impose une surveillance permanente de la Toile, irriguée par la publicité "made in Google". Et cette surveillance est d'autant plus cruciale que le moindre doute sur la réalité de ces clics publicitaires ruinerait la confiance que les annonceurs placent en Google. Et par conséquent ruinerait assez vite Google lui-même.

Car, à la fragilité technique – qui grandira à mesure que l'infrastructure informatique de l'entreprise grossira – s'y ajoute une autre : celle du modèle économique. "Aujourd'hui, nous avons tendance à croire que Google est immortel. Mais il faut garder en tête que son modèle économique ne repose que sur la publicité ciblée, relève Bernard Benhamou, délégué interministériel aux usages de l'Internet. Or, la crise qui vient va lourdement peser sur les budgets publicitaires. Et au contraire de Microsoft, qui profite directement de la vente d'ordinateurs pourvus d'un système d'exploitation, Google ne dispose d'aucune rente de situation."

Pour l'heure, pas de risque. D'abord, Google ne dépend pas des gros budgets publicitaires mais plutôt de l'addition de ceux de millions de petits commerces en ligne et de sites qui diffusent des annonces pour augmenter leur trafic. Ensuite, la "prime au leader" du marché publicitaire en ligne implique que ses concurrents soient d'abord plus durement touchés. Au troisième trimestre 2008, Google a ainsi gagné 2 % de parts de marché sur ses concurrents, selon l'agence Efficient Frontier. "Les prochains mois diront si son modèle économique protège l'entreprise d'une crise d'ampleur", estime M. Benhamou, pour qui il est hasardeux de trop se reposer sur Google, "en particulier pour ce qui touche au patrimoine culturel". Principalement aux livres numérisés et stockés sur les serveurs de l'entreprise californienne.

Pour la bonne conservation et la garantie d'un accès rapide aux données stockées, la

santé économique de l'entreprise est cruciale. Car Google est moins souple que ne le laisse supposer sa partie visible sur le Web. L'entreprise doit gérer et assurer la maintenance d'un outil industriel aux dimensions pharaoniques. Chaque centre de données est un entrepôt, souvent gigantesque, qu'il faut installer non loin d'une source d'énergie – un barrage ou une centrale nucléaire – et qu'il faut également refroidir… Ces questions sont si prégnantes et si problématiques que la société a déposé, récemment, un brevet sur un système lui permettant d'installer des centres de données offshore, utilisant l'eau de la mer pour le refroidissement et la houle pour l'alimentation énergétique ! A mesure que l'entreprise se développe, à mesure qu'elle grandit et que son cloud grossit et se complexifie, ses charges fixes augmentent et sa profitabilité devient plus impérieuse…

Croissance ad libitum

Or la règle du jeu, en économie de marché, est le développement perpétuel. La croissance ad libitum. Google l'a mieux compris que quiconque. Son essor est une histoire en trois actes. Acte I : l'entreprise indexe et organise l'espace public directement accessible à ses algorithmes, c'est-à-dire principalement le Web. Acte II : la société entreprend d'indexer et de conserver des éléments de la sphère privée, c'est-à-dire courrier, documents divers, etc. Acte III : Google construit l'information qui n'existe pas pour l'indexer et la rendre searchable – ou "recherchable". Google s'offre un satellite pour améliorer Google Earth. Google parcourt et photographie toutes les rues des grandes villes du monde pour proposer Google Street View. Google investit dans une société qui offre un service d'analyse du génome (lire l'encadré ci-contre)…

Google fabrique, pour l'indexer, l'information qui n'existe pas encore. Des grands espaces de la géographie terrestre aux réarrangements infinitésimaux des nucléotides de notre patrimoine génétique, le spectre des données conservées et compilées par Google semble devoir s'étendre indéfiniment. Une "bulle" dont on pourrait craindre qu'elle n'éclate un jour : la chose est à la mode.

Stéphane Foucart Publicité, mode d'emploi

Google vit de la publicité, qui représente 99 % de ses immenses revenus – 16,593 milliards de dollars de chiffre d'affaires en 2007. Pour générer de tels gains, l'indiscrétion est-elle le mal nécessaire de la réclame en ligne telle que l'entreprise américaine la pratique ? Selon Peter Fleisher, responsable de la protection des données personnelles chez Google, l'entreprise ne constitue pas des profils des utilisateurs pour leur proposer des annonces ciblées. La publicité est seulement, dit M. Fleisher, "contextuelle". Qu'est-ce que cela signifie ? Chez Google, la publicité revêt de plusieurs aspects. Il peut s'agir d'annonces générées par l'entrée d'un ou plusieurs mots-clés dans le moteur de recherche de la société. Une recherche sur tel ou tel type d'automobiles produira des liens publicitaires aboutissant, par exemple, à des sites Web de vente de ces véhicules… La publication d'annonces (Google Ads) sur un site Web obéit à un autre modus operandi. Lorsqu'un annonceur décide de faire appel à Google, la société procède à une analyse de l'ensemble des pages dont les espaces publicitaires lui ont été confiés. Ses algorithmes repèrent les pages les plus propices pour accueillir tel ou tel lien publicitaire. Google sait automatiquement que l'annonce d'un fabricant d'automobiles est plus à sa place sur le blog d'un passionné d'automobile que sur le site d'une fondation luttant contre le changement climatique… A chaque fois qu'un internaute clique sur l'un de ces liens publicitaires, l'annonceur règle une petite somme (en général quelques centimes d'euros) à Google. Qui en reverse une part au propriétaire du site-support de la publicité. Autre manière d'écouler la pub : l'analyse du courrier acheminé et affiché via Gmail, la

messagerie en ligne de Google. A chaque fois qu'un courriel est lu sur cette interface, la société en analyse automatiquement le contenu et y accole des liens publicitaires adéquats. Ainsi, si vous recevez sur votre compte de courrier Gmail un message d'un proche vous demandant quelle montre vous préférez pour votre anniversaire, il y a de fortes chances pour que des liens publicitaires d'horlogers-bijoutiers apparaissent sur le côté de l'écran.

Dix ans d'expansion

1998 Larry Page et Sergey Brin, 25 ans, fondent Google dans un garage de Memlo Park (Californie). A l'origine de la société, un moteur de recherche d'un nouveau genre conçu pendant leurs études à l'université de Stanford. 1999 En février, Google.com gère 500 000 requêtes par jour. En août, 3 millions. 2000 Le moteur de recherche est désormais disponible en onze langues, dont le français. Google Adword, programme de liens publicitaires, compte 350 clients à son lancement. 2001 Image Search donne accès à 250 millions d'images. 2002 Google News, une revue de presse automatisée, est lancée. La version francophone du service sera attaquée en justice par la presse belge et par l'AFP pour violation de copyright. 2003 Google annonce un projet de numérisation et d'indexation de livres : Google Print, futur Google Book Search. Inquiétudes aux Etats-Unis et en Europe sur les questions de droit d'auteur, tollé à la BNF, qui met en place un projet européen concurrent. 2004 Le siège mondial s'installe au "Googleplex", à Mountain View (Californie). Entrée en Bourse le 18 août ; l'action, à 85 $, atteindra 700 $ fin octobre 2007. Google lance la messagerie en ligne Gmail. 2005 Google Earth diffuse des images satellites de la terre entière. 2006 Google rachète le site de vidéos en ligne YouTube. 2007 Google obtient l'autorisation de l'autorité américaine de la concurrence pour le rachat de la régie de publicité en ligne DoubleClick. 2008 Google capte 40 % des dépenses de communication sur Internet. Le 2 septembre, lancement de Chrome, un navigateur Internet. Le 22 octobre, lancement aux Etats-Unis de G1, le smartphone de Google.

L'analyse d'ADN par alliance

Google, bientôt banque de données génétiques ? Ce n'est pas impossible. La société de Sergey Brin et Larry Page a récemment investi dans une start-up, 23andme.com, fondée par l'épouse de Sergey Brin, qui propose à ses clients de passer au crible leur ADN pour une somme relativement modique (moins de 400 dollars). Via son site Web, la société donne ensuite accès aux prédispositions à certaines maladies, fournit des informations sur l'ancestralité en fonction des caractéristiques de certains chromosomes ou de l'ADN mitochondrial… Sergey Brin a même donné un coup de pouce médiatique à l'entreprise de son épouse en ouvrant un blog dont le premier (et unique) billet, publié le 18 septembre, raconte comment il s'est rendu compte, grâce à 23andme.com, qu'il était porteur d'une mutation génétique le prédisposant à la maladie de Parkinson.