A la recherche de la "mémoire" du web : sédiments, traces et temporalités des documents en ligne

Olivier Ertzscheid, Gabriel Gallezot, Brigitte Simonnot

Ce texte est la version auteur de

Ertzscheid O., Gallezot G., Simonnot B. (2013) À la recherche de la « mémoire » du web : sédiments, traces et temporalités des documents en ligne. p53-68 Dans C. Barats (dir), *Manuel d'analyse du web*. Paris : Armand Colin. ISBN : 978-2-200-28145-8

À la recherche de la « mémoire » du web : sédiments, traces et temporalités des documents en ligne

Olivier Ertzscheid
Université de Nantes. IUT de la Roche-sur Yon

Gabriel Gallezot
I3M, Urfist, Université de Nice Sophia Antipolis

Brigitte Simonnot
CREM, Université de Lorraine

Introduction

Les projets de Vannevar Bush (1945) préfiguraient le réseau internet comme une extension de notre mémoire (memex). Si le web des débuts, conçu initialement comme un dispositif universel de partage et d'annotation de documents entre chercheurs, était davantage axé sur l'accès à l'information et la communication que sur l'archivage pérenne, force est de constater que le web actuel est devenu presque hypermnésique : le taux de rappel doit y devenir optimal. Le mouvement des textes publiés en ligne (déplacements ou suppressions) a longtemps été et reste encore un obstacle à la référence. À présent, des projets tentent de pallier cette instabilité, d'abord à l'initiative d'associations telle que la fondation Internet Archive[1] ou plus récemment par des institutions, comme la constitution par les grandes bibliothèques nationales d'un dépôt légal du web[2]. Le glissement terminologique opéré par Tim Berners-Lee du « web sémantique » au « web de données », plonge le web et ses usages dans une réalité plus prosaïque. D'autres théoriciens du web évoquent encore *The Stream*

[1] Fondation américaine qui s'est donné pour objectif l'archivage d'un grand nombre de sites web et de matériaux documentaires libres de droits (films, textes, etc.). Voir www.archive.org.

[2] Pour la France, *Décret n° 2011-1904 du 19 décembre 2011 relatif au dépôt légal*, JORF n°0295 du 21 décembre 2011 page 21638, texte n°42. http://www.legifrance.gouv.fr/

(Spivack, 2009) pour signifier la publication en « temps réel » des informations ou soulignent les effets de dispositifs socio-techniques comme dans *machine is us/ing us* (Wesh, 2007). La question de la trace, de la traçabilité, et plus largement des différents types d'engrammation[3] possibles et de leurs objectifs est au cœur même du développement de la Toile de l'internet, dans ses outils comme dans ses usages. Si l'évolution du web donne lieu à de nombreux discours et prospectives, nous souhaitons analyser cette évolution et ses effets possibles sur les pratiques de recherche en sciences humaines et sociales (SHS). Pour cela, nous distinguons le web documentaire, avec les sources qui relèvent de la discipline du chercheur en SHS, et celles qui peuvent devenir son terrain d'investigation ou d'expérimentation, avec la multiplication des traces « sociales » (indices, empreintes, documents, écritures) susceptibles de révéler des pratiques. Le web s'est construit sur trois principes : des contenus bénéficiant d'un adressage stable (URL), librement accessibles et explicitement qualifiables au moyen des liens hypertextes. Ces trois piliers sont aujourd'hui ouvertement menacés. Les logiques d'inscription de traces sont en effet influencées par les industries de la recommandation, faisant peser le risque de passer du rêve d'une mémoire sans contrainte à une mémoire sous contrainte.

1 Le web : de la mémoire aux traces

Toute civilisation repose sur une mémoire, un patrimoine commun fait d'ensembles de traces sélectionnées et de souvenirs réactivés. Parmi ces objets de mémoire, le document occupe une place non exclusive mais singulière. Le terme « document » désigne un ensemble, tout à la fois support d'information, données enregistrées sur ce support et leur signification, « servant à la consultation, l'étude, la preuve ou la trace», considéré comme une unité autonome (Boulogne, 2004 : 80). Représentation présente d'une chose absente (un événement, une réflexion, un échange), le document organise matériellement des marques selon des procédés

[3] Par engrammation, nous entendons la mémorisation par écriture d'un flux informationnel.

socialement convenus. L'environnement numérique, en bouleversant sa temporalité et sa granularité, invite à revisiter les approches des processus d'information et de communication, qu'ils soient analysés sous l'angle technique, sémiotique, pragmatique ou sociologique, pour les penser en termes « d'empreintes, de signatures et de traces » (Merzeau, 2009).

1.1Traces explicites, traces implicites

Comment définir la notion de trace telle qu'elle est actualisée par l'informatique et plus particulièrement par le web et ses applications ? Nous pouvons opérer une première distinction entre traces explicites et implicites. Les premières sont constituées des écrits et productions diverses (textes, images, vidéos enregistrés, données) publiés en ligne qui relèvent d'une volonté exprès de diffuser et qui, le plus souvent, peuvent être modifiés ou supprimés après publication. C'est le cas des écrits de toutes sortes, des billets de blogs aux tweets, voire des étiquetages par mots-clés (*tags*) dont l'internaute sait qu'ils seront consultables par d'autres et qui correspondent à des extériorisations délibérées, avec une volonté plus ou moins affirmée de transmettre dans l'espace et le temps. Mais le web, ce sont aussi les traces implicites, prélevées souvent à l'insu de l'internaute lors de ses interactions en ligne. Dans ce cas, le terme « trace » désigne une marque laissée par l'utilisateur d'une application informatique, témoignant d'un contact passé avec cette application qui en programme l'inscription. La superposition des couches applicatives qui caractérise l'architecture du web démultiplie leur accumulation selon des logiques diverses. Entre explicite et implicite, certains dispositifs de captation de traces ont un statut hybride, comme les boutons « j'aime » (Facebook)[4] ou « +1 » (Google) qui sollicitent une action délibérée mais que l'internaute ne pourra effacer. Comme le rappelle Jean-Michel Salaün (2006), « le web favorise conjointement deux mouvements opposés : le développement d'échanges

[4] Voir dans cet ouvrage : Étienne Candel et Gustavo Gomez-Meija, « Signes passeurs et signes du web : le bouton *like*, ou les ressorts d'un clic ».

spontanés (conversations) et leur fixation sur un support public, pérenne et documenté. Autrement dit, le web transforme automatiquement ce qui relevait de l'intime et de l'éphémère en document ou en proto-document ».

1.2D'un protocole sans mémoire à la collecte systématique

Dans le cas des navigations web, le protocole initial qui permet d'accéder aux documents – le protocole http – est dit « sans mémoire » au sens où chaque nouvelle demande à un serveur web n'est pas interprétée en lien avec la précédente. Cependant, les serveurs conservent des historiques de transactions, sortes de journaux des requêtes qu'ils reçoivent (les *logs*). Ces journaux d'événements, d'abord conçus pour permettre aux informaticiens de gérer d'éventuelles erreurs, constituent la première collecte systématique de traces externalisées dans l'accès au web. Les données ainsi collectées, pour rudimentaires qu'elles soient, alimentent des analyses de transactions web, des statistiques d'accès aux études des comportements des internautes : qui vient sur le site, que vient-il chercher ? Sous couvert de mieux répondre aux « besoins » des utilisateurs, ces données sont disséquées, filtrées, interprétées. Les navigateurs collectent aussi les traces de navigation de l'internaute, cette fois mémorisées sur son ordinateur (l'historique, le cache, les « cookies »), mais accessibles aux applications extérieures. Ces données sont désormais exploitées, par exemple, par les moteurs de recherche pour « personnaliser » les résultats (Simonnot, 2012). Chaque application organise sa propre collecte de fragments et conserve en mémoire des traces de nos actions.

Comme la mémoire humaine, les mémoires informatiques sélectionnent les événements dont elles conserveront la trace à travers les algorithmes qui permettent de les renseigner. Toute communication médiatisée par l'ordinateur, y compris par l'internet, est susceptible de faire l'objet d'un enregistrement, d'une inscription : des documents publiés aux billets de blogs, des échanges sur les forums de discussion ou via les réseaux sociaux aux discussions en temps réel (*chat*). L'internet est un dispositif qui, par la rapidité qu'il propose, rapproche les

communications de l'oralité mais en même temps, ses applications permettent d'inscrire de manière plus ou moins pérenne les traces de ces échanges dans les mémoires informatiques. Les concepteurs d'applications décident ou non de mettre à disposition du public ces traces sous différentes formes[5]. Cette externalisation généralisée de données sur nos actions font qu'elles nous échappent : coupées de leur contexte, elles sont agrégées à d'autres et interprétées avant de nous être reproposées pour orienter nos parcours.

1.3Analyse des traces web en SHS

Pour les chercheurs en sciences humaines et sociales, ces traces collectées peuvent apparaître comme une nouvelle manne pour analyser pratiques et comportements via des dispositifs numériques interposés. La tentation est forte de considérer la collecte automatique des traces comme une solution aux potentielles « pertes d'objectivité » connues des chercheurs qui procèdent à des entretiens, comme un moyen d'atteindre l'indicible. Ce serait oublier que la captation de ces traces a forcément été pensée en amont, elle est intégrée à des applications qui filtrent et donnent une forme préconçue aux actions qu'elles rendent possibles. L'analyse des traces numériques permet surtout de comprendre comment les usagers pratiquent avec tel ou tel dispositif ou tel logiciel, selon un cadre préconçu. Le procédé a été employé à de nombreuses reprises dans le domaine de la recherche d'information, avec l'analyse des historiques de transactions de moteurs de recherche. Dans ce domaine, les chercheurs ont montré que les résultats obtenus pour un moteur donné ne peuvent pas être automatiquement transférés à un autre (Jansen & Spink, 2006), ce qui atteste du caractère situé du traçage numérique.

Le recours aux traces pourrait relever d'un « paradigme indiciaire » (Ginzburg, 1989 ; Serres, 2002) : il s'agit de prêter attention au moindre détail, à toute singularité plutôt que de chercher

[5] Par exemple, Yahoo Search! affiche en page d'accueil les mots clés les plus demandés dans les recherches d'information, Google Trends (http://www.google.fr/trends/) et Google Insights (http://www.google.com/insights/search/?hl=fr) permettent de consulter les « tendances » des requêtes des internautes sur le moteur, Twitter affiche également les mots ou expressions les plus courants.

des généralités, de procéder plutôt à une analyse qualitative que quantitative. C'est de ce paradigme que relèvent les pratiques de « personnalisation » : à chaque internaute ses résultats. Toutefois, l'exploitation des traces implicites s'éloigne du paradigme indiciaire dans d'autres exploitations marketing, lorsqu'il s'agit de dégager des tendances pour les marchés. Un tel dispositif de traçage numérique pourrait ainsi être considéré comme « un dispositif d'observation "total" qui promet de révolutionner les méthodologies classiques en effaçant le clivage entre micro et macro, entre qualitatif et quantitatif » (Rieder, 2010). Il serait susceptible d'éliminer les « biais » connus des chercheurs adeptes des méthodes qualitatives (notamment entretiens et récits de vie) où la personne interrogée peut être tentée de composer un récit plausible à destination du chercheur. Ce mythe de la « vérité » oublie que, d'une part, ces traces reflètent d'abord les critères sélectionnés par ceux qui en organisent la collecte et, d'autre part, ne fournissent pas d'éléments contextuels suffisants par eux-mêmes : elles indiquent surtout ce que l'internaute a fait avec telle ou telle application, en aucun cas ce qu'il voulait faire et les autres moyens qu'il a mis en œuvre pour arriver à ses fins. Ensuite, leur « authenticité » peut être mise en doute, comme en attestent certains épisodes dans l'histoire des moteurs de recherche (Google-bombing, référencement) ou du marketing en ligne (fraude au clic). Penser que ces traces sont « authentiques », c'est tomber dans l'idéologie, celle du marché dont la liberté serait garante de la vérité des prix (Rieder, 2010). C'est oublier la manière dont les applications qui collectent ces traces contraignent et souvent interviennent dans les pratiques des internautes. L'illusion de l'observateur « invisible » mérite d'être interrogée, aussi bien dans l'exploitation des traces implicites que dans les analyses des échanges via forums de discussion ou réseaux sociaux interposés, cette interrogation renvoyant d'abord à l'éthique du chercheur.

2 Le web comme ressource pour le chercheur

Le web de flux impose son rythme : celui de l'immédiateté, du temps réel, du renouvellement

permanent. Comment concilier la temporalité de ces nouveaux dispositifs avec celle d'un travail borné par les impératifs d'une recherche scientifique ? Les flux passent, sont mixés et se sédimentent mais les fouilles sont rares et difficiles. La question de l'accès pérenne aux contenus et aux liens pose problème pour les appréhender « en contexte ». Le web est en quelque sorte passé d'un « web de stock », où documents et données étaient stockés sur un serveur et « figés » par une URL, à un « web de flux » où l'information circule, est détachée de son support et de sa forme initiale. Ce glissement est notamment sous-tendu par le passage « HTML/documents » à « XML-RDF[6]/datas ». Le web supporte une information versatile qui se multiplie et s'incarne dans différentes formes de documents (fixe/dynamique, long/court, textuel/hypermédia, etc.) stockés dans « les nuages ». Avec la multitude et la duplicité des formes de document, certains en appellent à nouveau à la curation (sélection, tri, édition et partage des flux d'information sur un thème).

2.1Le web documentaire : accès facilité à la production scientifique

Dans le célèbre « *As we may think* », Vannevar Bush (1945) propose de relier l'ensemble des connaissances entre elles pour naviguer au gré de nos schèmes cognitifs. Le web dans sa première version, tel qu'il est conçu au CERN[7], se cale sur cette vision de l'échange scientifique. Si le web commercial est venu bouleverser ce projet à partir de 1994, l'apparition de ressources dédiées à la science permet de considérer un certain continuum avec le « memex » et renouer avec les origines du web et de la publication scientifique.

Des échanges épistolaires à la naissance des premières revues (1665, *Journal des Scavans* et *Philosophical Transaction*), jusqu'à l'apparition des bases de données (1960-70), pendant plus de 300 ans, le contenu intrinsèque de chaque texte a été l'unité de référence. Les bases de données ont d'abord favorisé le développement des projets bibliographiques (SCI, *science*

[6] RDF ou *Resource Description Framework* est un standard d'échange de données sur le web, qui permet de spécifier des triplets exprimant des relations entre des ressources de nature diverse (http://www.w3.org/RDF/).

[7] C'est au CERN (organisation pour la recherche nucléaire) que Tim Berners-Lee et Roger Caillau ont inventé la Toile ou Word Wide Web en 1990, le premier serveur web ayant vu le jour en 1991 en Californie (Source : http://public.web.cern.ch/public/fr/About/WebStory-fr.html).

citation index), puis elles ont renforcé la collecte des données issues des terrains scientifiques et plus tard, par le biais de CGI (*Common Gateway Interface*), la publication de ces mêmes données sur le web. La gestion de contenu telle que nous la connaissons aujourd'hui sur le web ne peut se concevoir sans base de données. Du site web au blog en passant par les moteurs, les wikis ou les archives ouvertes, les cyberinfrastructures s'appuient essentiellement sur une plate-forme étayée par une base de données. C'est le principe de découpage de l'information en unités informationnelles ordonnées par un schéma conceptuel qui autorise une manipulation granulaire du contenu.

Les bases de données bibliographiques sont la trace des publications scientifiques, elles jouent un rôle essentiel de mémorisation du contexte de la production textuelle scientifique. Il convient de noter que toutes les publications scientifiques ne sont pas référencées par ces bases. Même s'il y a un grand nombre de bases[8], il y aussi un nombre de publications toujours plus important[9], non seulement en raison de l'évolution de la science et des ramifications disciplinaires mais aussi de la fièvre de l'évaluation[10] qui fait « monter les enchères ». Des acteurs commerciaux et institutionnels maintiennent ces bases bibliographiques. On relèvera que les acteurs commerciaux (Thomson-Reuters, Elsevier, Ebsco, etc.), les éditeurs et agrégateurs, proposent via la référence bibliographique souvent l'accès (payant) au texte intégral, quand les acteurs institutionnels (ABES, INIST,...) « chassent » les références bibliographiques. L'évaluation de la recherche s'effectue en partie (de manière importante dans les sciences et techniques, moindre en sciences humaines et sociales) essentiellement sur une base d'un éditeur commercial (Thomson-Reuters avec le *Web of Science* - WoS) et sur une

[8] Le Gale Directory of Databases recense plus de 20 000 banques de données. http://www.gale.cengage.com/

[9] À titre d'exemple, selon GFII, 23 000 revues scientifiques à comité de lecture étaient publiées dans le monde en 2010, dans lesquelles paraissent annuellement 1,4 million d'articles. Le nombre d'articles publiés par an et le nombre de revues ont augmenté régulièrement pendant deux siècles, respectivement de 3 % et 3,5 %. http://www.gfii.fr/fr/document/groupe-de-travail-gfii-sur-le-libre-acces-mise-en-ligne-de-la-synthese-des-discussions-et-des-recommandations

[10] Dossier « La fièvre de l'évaluation », *Revue d'histoire moderne et contemporaine* 5/2008 (n° 55-4bis), p. 5-6.

forme d'écrit scientifique : l'article de revue. La place occupée par de telles sociétés commerciales dans le domaine des bases bibliographiques et le modèle sous-jacent (re)posent la question d'une mémoire publique et librement accessible, à laquelle l'Open Access et le web 2.0 donnent des éléments de réponse.

2.2Archivage scientifique et partage de références : mémorisation publique de la science

Archiver le web a posteriori pose des défis d'ampleur. Au niveau mondial, on pourra citer Internet Archive, dont le projet de départ était d'archiver tout le web et, plus récemment en France, le Centre Informatique National de l'Enseignement Supérieur (CINES) qui assure l'archivage de données scientifiques, patrimoniales ou administratives.

L'Open Access est un mouvement qui prône le libre accès aux résultats de la recherche. Il favorise ainsi une mémorisation publique de la science et concoure à la constitution d'un patrimoine (textuel) scientifique mondial. Il se développe selon deux axes : la « voie verte », c'est-à-dire l'auto-archivage par les chercheurs dans les archives ouvertes (AO), et la « voie dorée » qui concerne la publication dans des revues en libre accès (RLA). Pour les RLA, il s'agit d'inciter les communautés savantes et éditeurs publics à publier eux-mêmes et rendre accessibles leurs publications (articles, parfois aussi collections), renouant en cela avec le principe des premières revues. Pour les AO, il s'agit d'inciter les chercheurs à auto-archiver leurs textes (*pre-print* ou *post-print*) sur des plateformes interopérables (grâce à un protocole nommé OAI-PMH) et qui peuvent ainsi échanger entre elles des notices. Des moissonneurs OAI récoltent l'ensemble des notices et proposent alors une forme de base bibliographique globale, renouant ainsi avec le souhait de mémoriser les traces (les références bibliographiques) des écrits scientifiques et proposant du même coup l'accès au texte intégral. La plateforme Isidore[11] ouverte en 2010 à l'initiative du très grand équipement Adonis du

[11] http://www.rechercheisidore.fr

CNRS est une alternative publique pour les SHS aux projets d'entreprises commerciales de type Google Scholar, Microsoft Academic Search ou Scirus[12].

Les outils du web 2.0 ont leur « spéciation » scientifique. Sous ce vocable nous rassemblons les plates-formes de partage de références, les weblogs et les plates-formes de réseau social à caractère scientifique. Les plates-formes de partage de références bibliographiques ou webographiques (par exemple CiteULike, Delicious, Diigo, Zotero, Mendeley) sont intéressantes à observer en ce qu'elles s'apparentent au Memex et aux bases de données bibliographiques. Elles mémorisent et organisent les références, elles les lient. Elles permettent la mise en commun de traces et un travail collaboratif sur ces dernières. Les weblogs représentent la version « en ligne » des carnets de recherche. Ils offrent à lire les traces du cheminement scientifique, la mémorisation du travail de terrain. Du site de chercheur aux « agrégateurs » de billets (Postgenomic), en passant par les plates-formes dédiées à la recherche (Hypotheses.org), les blogs de chercheurs occupent désormais une place importante dans l'agora scientifique. Les plates-formes de réseau social comme ResarchGate ou SciLink[13] agrègent un certain nombre de traces liées à chacun des membres du réseau (bibliographies, textes, appartenance à un laboratoire ou université) et façonnent de nouvelles relations. Deux formes de réseaux co-existent : ceux partant du groupe pour partager des informations, d'autres s'étayant sur les informations partagées pour construire des communautés (Gallezot & Le Deuff, 2009).

2.3Le web de données

Le web ne se résume pas seulement aux bases bibliographiques et entrepôts de documents. Il organise aussi des bases de données factuelles qui sont la mémoire des résultats d'expérimentations, de recueils, d'enquêtes scientifiques. Ces bases peuvent être d'accès

[12] http://scholar.google.fr/, http://academic.research.microsoft.com/, http://www.scirus.com/

[13] http://www.researchgate.net/, http://www.scilinks.org/ ou encore des outils « intranet » comme BibApp http://bibapp.org/

restreint à l'usage exclusif d'une communauté, d'un laboratoire, ou à accès public. On peut citer par exemple les bases génomiques qui recueillent le patrimoine génétique de différentes espèces, des bactéries à l'homme. Les internautes peuvent aussi être directement associés aux collectes : la plateforme patientslikeme.com par exemple, permet aux patients de partager leur expérience autour d'une maladie, d'un traitement médical et de décrire leurs symptômes (Arnott Smith & Wicks, 2008). Dans le domaine de la recherche, les cyberinfrastructures[14] agrègent, gèrent, traitent des sommes de documents mais aussi de données pour permettre aux chercheurs de travailler sur des informations de qualité, de manière pérenne. Elles se présentent sous la forme de plateformes agrégatrices de traces, de textes, de données et proposent des outils de traitement de ces informations. Ici c'est le quatrième paradigme de Jim Gray (Hey et als, 2009), l'aspect computationnel de la recherche qui est mis en avant. La recherche scientifique se fait par un traitement informatique de données collectées et mémorisées par ailleurs, ce qui permet d'interroger sous un angle nouveau les conditions selon lesquelles elles sont constituées. La conception des données recueillies est toujours, en effet, le fruit en amont d'un travail d'interprétation. Ce dernier est-il toujours explicité ? Il ne suffit pas que des données soient dans un format apparemment similaire : croiser des données de sources diverses mérite réflexion et précautions, sans oublier l'éthique qui doit y présider.

L'entropie, l'apparent « désordre » initial du web a, dans un premier temps, progressivement fait place à une hiérarchisation organique, mise en mémoire et en accès par les annuaires et les moteurs de recherche. L'essor du web contributif (souvent résumé par l'expression web 2.0) a ensuite permis de dépasser et de transgresser littéralement cette hiérarchisation monolithique en y réinjectant de l'humain, du social, de l'aléatoire et en venant bousculer les ingénieries de l'accès mises en place jusqu'alors, notamment celle de Google et du principe de

[14] Quelques exemples de cyberinfrastructures : la plateforme ISIDORE du TGE ADONIS (http://www.rechercheisidore.fr/), OpenAire (http://www.openaire.eu/), SIDR (http://www.sidr-isb.eu/), Grid5000 en informatique(https//www.grid5000.fr/) ou encore l'initiative européenne DARIAH (http://www.dariah.eu/) pour les humanités et l'art.

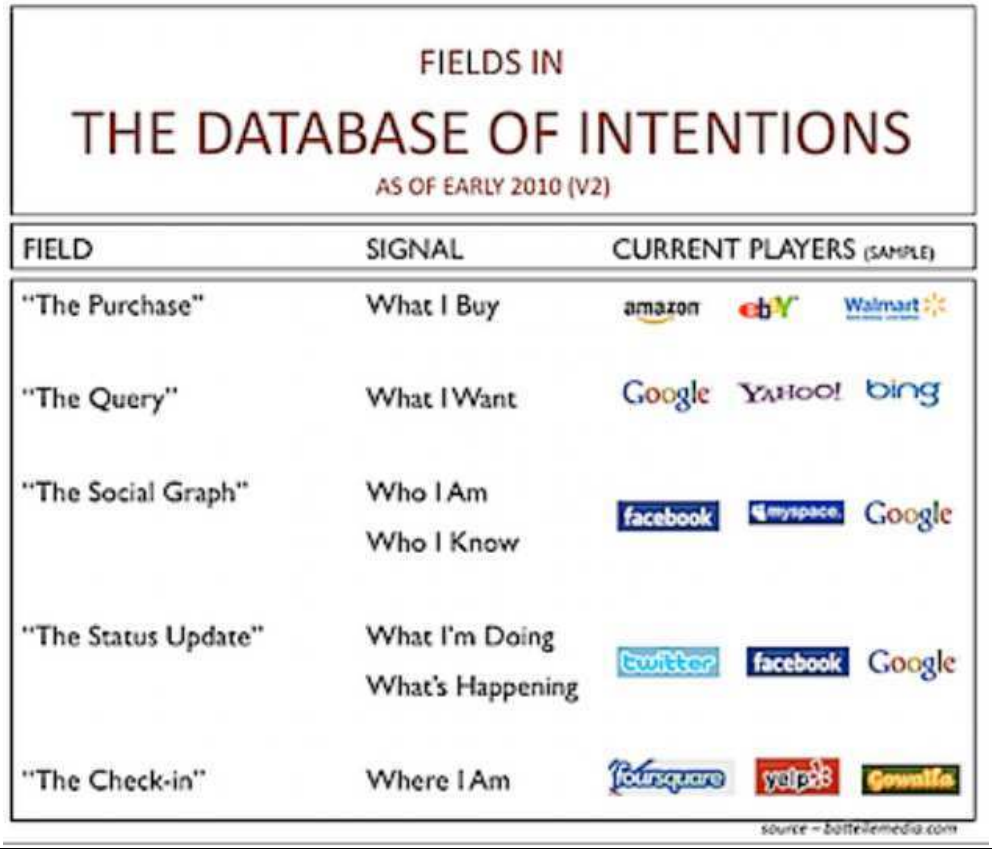
« popularité » qui sous-tend l'algorithme du PageRank. Ce second temps a signé également l'entrée en masse d'unités documentaires privées ou intimes dans le radar mémoriel assuré par l'ingénierie des différents moteurs de recherche et ce, notamment avec l'explosion et la banalisation de systèmes de publications presqu'exempts de toute barrière technique, les blogs. L'accessibilité des données, leur caractère « public », n'autorise pas pour autant le chercheur à passer outre l'indispensable éthique pour en penser les exploitations possibles (Boyd & Crawford, 2011).

3 Les industries de la recommandation : un rapport à la mémoire par délégation

Le couplage structurel entre une ingénierie documentaire dédiée au repérage et à l'accès d'une part, et la granularité toujours plus dense et fragmentaire de traces documentaires personnelles d'autre part, en se conjuguant à une libération des pratiques et des comportements d'achat connectés, a donné naissance aux industries de la recommandation. Elles furent les premières à modéliser et à systématiser l'articulation entre des logiques commerciales courantes et des stratégies de personnalisation. La collecte systématique des traces, comportements et intentions attachés à un profil donné permet de les comparer avec ces mêmes traces, comportements et intentions cette fois déclinés à l'aune de moyennes statistiques portant sur des communautés très étendues (les 850 millions de membres de Facebook, l'ensemble des clients d'Amazon, l'ensemble des requêtes déposées sur Google, etc.). Ce qu'un analyste américain (Battelle, 2010) baptisa du nom de "base de données des intentions", chacune ayant son domaine de prédilection (figure 1).

Pour exister et fonctionner, ces industries de la recommandation ont besoin de mobiliser des ingénieries mémorielles spécifiques autorisant les différents couplages décrits précédemment. Elles peuvent alors maintenir vivace l'illusion de pouvoir prédire certains comportements encore au stade intentionnel quand il ne s'agit pourtant très trivialement que d'exploiter systématiquement une gigantesque base de données relationnelles. Aujourd'hui, l'avènement

de l'informatique en nuage, du *cloud computing*, aussi bien dans les usages courants du web (archivage et stockage de matériaux documentaires personnels) que dans les stratégies commerciales de petites entreprises ou de grands groupes (Saas, *Software as a service*) est la conséquence logique d'un rapport à la mémoire de plus en plus vécu par délégation, sinon par procuration. L'externalisation de nos mémoires documentaires, privées ou intimes, est devenue une caractéristique marquante de nos pratiques informationnelles connectées, avec le risque de se trouver « privé » desdites mémoires au gré des fluctuations du cours de bourse de l'un des acteurs du Cloud ou du changement des conditions générales d'utilisation (CGU) d'un service donné.



Domaine de spécialité	Signale	Quelques acteurs
« la Requête »	Ce que je veux	Google Yahoo Bing
« le Graphe social »	Qui je suis Qui je connais	Facebook Myspace Google
« La mise à jour de statut »	Ce que je fais Ce qui arrive	Twitter Facebook Google
« L'enregistrement »	Où je suis	Foursquare Yelp Gowalla

Figure 1 - La base de données de nos intentions début 2010 (d'après John Battelle, batellemedia.com)

L'économie de la recommandation est aussi une économie de la saturation. Les « like » et autres "+1", les stratégies du graphe social éparpillées sur un ensemble de sites de plus en plus large, menacent chaque jour davantage l'écosystème du web parce que l'activité de « pousse bouton » se substitue à celle de l'établissement de liens hypertextes. Plutôt que d'instancier des documents en les décrivant au moyen de liens (rappelons ici que Google n'auraient jamais pu exister sans l'œuvre quotidienne et bénévole de millions d'internautes, indexeurs sans le savoir) nous préférons les signaler, les "liker", sans en retirer aucun gain cognitif individuel ou collectif, et en déléguant la gestion de ces signalements éparpillés à des sociétés tierces sans se questionner sur ce que peut valoir pour tous un signalement non pérenne, un signal éphémère (Ertzscheid, 2010).

Les premiers textes sur le web sémantique décrivaient déjà l'évolution d'un web consistant principalement en documents destinés à la lecture humaine en une infrastructure incluant données et informations que les ordinateurs peuvent manipuler, des informations actionnables. La mémorisation numérique procède par activation. Les moteurs eux-mêmes sont utilisés comme des marque-pages de nos navigations, des supplétifs de notre capacité à retrouver et donc à se souvenir, ils sont – au sens propre comme au sens figuré – jugés et estimés sur leur « taux de rappel »[15]. Ces pourvoyeurs de requêtes transactionnelles ont naturellement à voir avec l'élaboration de mémoires transactives (Sparrow *et al.*, 2011) qui modifient une composante essentielle de l'humain, à savoir la manière et les outils dont il dispose pour interagir avec sa mémoire, avec ses souvenirs, avec sa propre histoire. Mémoire des textes avec GoogleBooks, mémoire des conversations avec Facebook ou Twitter, mémoire de la presse avec Google News, mémoire photographique avec Flickr et FlickrCommons, mémoire topographique avec GoogleMaps, mémoire des sons avec SoundCloud, mémoires des signets et des notes avec Diigo et Delicious, etc. Autant de mémoires qui se mettent en place sur

[15] Le taux de rappel est défini par le nombre de documents pertinents retrouvés au regard du nombre total de documents pertinents disponibles. Avec le taux de précision, il est l'une des deux métriques fondamentales dans le domaine de la recherche d'information.

plusieurs types de territoires documentaires : territoires de la qualification (avec nos mémoires littéralement « documentaires »), territoires de socialisation (avec nos mémoires affectives, personnelles et sociales) sans oublier les territoires du marketing (avec nos mémoires « actionnables », intentionnelles : sorties au cinéma, restaurant, achats, déplacements, etc.)

Mais le cycle mémoriel numérique a ceci de radicalement différent qu'il procède d'une engrammation non pas choisie mais consubstantiellement contrainte. Que l'on interagisse par l'écrit, par l'image ou par la voix – et sauf à utiliser certains dispositifs dédiés nécessitant une maîtrise ou une connaissance technique particulière – il n'est pas possible de se soustraire à l'enregistrement par un tiers que l'on en est réduit à espérer être « de confiance ». L'équivalent d'une procédure d'*opt-out* – c'est-à-dire le principe d'une collecte par défaut des données des internautes, leur consentement étant considéré comme implicite – mise en avant par Google dès son arrivée sur le web est désormais étendu à l'ensemble des matériaux qui le constituent : « Nous nous souvenons de tout, à vous de n'avoir rien à vous reprocher ». Une dichotomie apparaît alors qui oppose deux grands types de rapports à la mémoire : celui de la collection et celui de l'itération. La collection relève d'une logique archivistique qui postule un cadre classificatoire initial à l'intérieur duquel il s'agit d'effectuer un certain nombre de choix, de mettre en place un certain nombre de filtres. Le numérique inaugure une nouvelle logique, algorithmique, qui est de l'ordre de l'itération : on laisse les choix se faire d'eux-mêmes à partir d'une formule, d'un programme initial sans cesse réitéré. Or, si le Pagerank a su faire la preuve de son efficacité pour organiser l'accès aux contenus, il a également souvent montré les limites d'un mode d'accès presqu'exclusivement fondé sur des métriques de popularité, mode d'accès qui constitue un risque pour le maintien d'une diversité éditoriale et qui, seul, est parfaitement impropre à inscrire des contenus dans un cadre de préservation et de mémorisation à moyen ou long terme.

4 Le web comme corpus ?

Dans le cadre ainsi décrit, comment constituer de nouveaux corpus et comment traiter le gigantisme de ceux mis à disposition ? Dans l'histoire des sciences, les scientifiques de tous les domaines, de toutes les époques, de toutes les disciplines, se sont en permanence efforcés de prendre l'ascendant sur leurs différents corpus ; pour pouvoir être exploitable, le corpus doit pouvoir être circonscrit par ceux qui prétendent en faire l'analyse. « *Il n'y a rien que l'homme soit capable de vraiment dominer : tout est tout de suite trop grand ou trop petit pour lui, trop mélangé ou composé de couches successives qui dissimulent au regard ce qu'il voudrait observer. Si ! Pourtant, une chose et une seule se domine du regard : c'est une feuille de papier étalée sur une table ou punaisée sur un mur. L'histoire des sciences et des techniques est pour une large part celle des ruses permettant d'amener le monde sur cette surface de papier. Alors, oui, l'esprit le domine et le voit. Rien ne peut se cacher, s'obscurcir, se dissimuler.*" (Latour, 1985).

Google Books, projet de numérisation lancé en 2005, dispose à ce jour de 4% de tous les livres publiés depuis deux siècles, en sept langues. Soit une estimation à hauteur de deux milliards de mots et 5,2 millions de livres numérisés (Cohen, 2010), tout simplement « le plus grand corpus linguistique de tous les temps » (Véronis, 2010). Autre type de corpus, celui de Facebook et de ses 850 millions de membres, soit le plus grand « corp(u)s social » numérique, le plus grand pan-catalogue des individualités et de leurs mémoires (Ertzscheid 2007, 2010b).

Du premier corpus, celui de Google, on ne pourra que se réjouir, pour ce qu'il représente de potentialités ouvertes dans l'aventure linguistique comme compréhension du monde, sous réserve qu'il soit et reste disponible aux chercheurs indépendants. Et l'on mettra du temps à en épuiser les possibles. Du second, on ne peut que continuer à raisonnablement s'alarmer devant des usages que seul contrôle et détermine le site hôte, compagnie cotée en bourse et devant satisfaire ses actionnaires.

L'informatique, les outils de la linguistique de corpus ont permis aux linguistes de rester les maîtres de corpus aux dimensions exponentielles. Même chose dans le domaine de la médecine avec la conquête du dernier grand corpus, celui du génome. Ces gigantesques corpus numériques, s'ils sont une formidable opportunité pour différents champs de recherche, posent aux scientifiques au moins deux questions cruciales : comment y avoir accès, et avec quelles règles, contraintes ou limites méthodologiques et éthiques. Et enfin comment s'assurer de la complétude ou de l'incomplétude dudit corpus quand celui-ci est entièrement géré ou mis à disposition par des sociétés régies avant tout par des logiques commerciales et non scientifiques. Individuellement comme collectivement, nous nourrissons en permanence des monstres calculatoires et industriels qui, dans certains domaines, sont en passe d'être les seuls capables de circonscrire des corpus qui relèvent pourtant du bien commun. Ce qui oblige à repenser totalement la question de l'archive et du rôle de la puissance publique dans la constitution, la gestion et l'accès à cette dernière. Du côté du document, cette engrammation permanente repose les questions qui fondent toute la science de l'archivistique dans une perspective foucaldienne, c'est-à-dire l'archive comme « la masse des choses dites dans une culture, conservées, valorisées, réutilisées, répétées et transformées. Bref toute cette masse verbale qui a été fabriquée par les hommes, investie dans leurs techniques et leurs institutions, et qui est tissée avec leur existence et leur histoire » (Foucault, 1968).

Conclusion : des alternatives au marché de la mémoire

Il est aujourd'hui frappant de voir comment s'inverse le mouvement initial : l'externalisation de mémoires documentaires, de nos mémoires « de travail », se renverse pour devenir une internalisation de parcours mémoriels intimes, reconstruits *a posteriori*, en fonction d'objectifs sur lesquels nous n'avons aucune prise et avec des modalités d'activation qui nous sont tantôt suggérées et tantôt imposées par les principaux vecteurs mémoriels numériques que sont les

outils de notre présence et de notre activité en ligne. Se pose alors la question triviale du devenir d'une société potentiellement hypermnésique, ou plus précisément, la question de la viabilité du devenir d'une société dans laquelle tout est à tout moment « retrouvable », « réaccessible », « réactivable ».

Pour cet oligopole de grands groupes commerciaux et de régies publicitaires qui dominent le marché de la mise en mémoire, l'enjeu principal est celui du contrôle de la constitution de cette mémoire, mais également du contrôle de la réaffectation possible d'une mémoire collective et d'agrégats de mémoires individuelles. Google, Facebook et les autres ont compris depuis longtemps que le contrôle de l'engrammation, de ce qui « fait mémoire », constituera pour eux la prochaine clé de leur suprématie, et donc de leur survie.

Préserver le web d'une évolution vers un marché de la mémoire nécessite des choix politiques forts et l'investissement des politiques publiques. Les sociétés humaines, les civilisations se construisent sur la mémoire. Sur une mémoire partagée et rassemblée et non sur des fragments mémoriels largement « partagés », en permanence « disséminés », épars. Comme il n'est pas de mémoire possible sans un véritable travail d'archive, l'archivage public des traces du web s'impose pour mettre en place des logiques différentes de celles instaurées par les acteurs commerciaux. Les traces peuvent aussi être mobilisées au sein d'applications qui les mettent au service de la réflexivité individuelle et de l'appropriation des dispositifs. Une condition pour que nos parcours et nos lectures soient véritablement créatifs et dépassent les lois de l'imitation.

Références bibliographiques

Arnott Smith, C., & Wicks, P. J. (2008). PatientsLikeMe: Consumer Health Vocabulary as a Folksonomy. *AMIA Annual Symposium Proceedings 2008,* (pp. 682–686).

Battelle J. (2010), "The Database of Intentions is Far Larger Than I Thought", *John Battelle's Searchblog*, 5 Mars 2010 http://battellemedia.com/archives/2010/03/the_database_of_intentions_is_far_larger_than_i_thought.php

Boulogne, A. (2004). *Vocabulaire de la documentation.* Paris, ADBS éditions.

Boyd, D., & Crawford, K. (2011). Six Provocations for Big Data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society,* September 21, 2011. Oxford Internet Institute: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431#122782.

Bush, V. (1945), "As We May Think", *The Atlantic Monthly*, 176 (1):101-108, Juillet 1945. <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/> (Traduction française disponible sur : http://mediateur.free.fr/web/hist_aswemaythink_fr.htm)

Cohen P. (2010), In 500 Billion Words, New Window on Culture, *New-York Times*, 16 décembre 2010 <http://www.nytimes.com/2010/12/17/books/17words.html>

Ertzscheid O. (2007), Les "nouveaux catalogues ou le catalogue en (r)évolution, *Arabesques*, n°48, Octobre - Novembre - Décembre 2007.

Ertzscheid O. (2010), « Le "like" tuera le lien », in *Affordance.info*, 16 Mai 2010 <http://affordance.typepad.com/mon_weblog/2010/05/le-like-tuera-le-lien.html>

Ertzscheid O. (2010b), "Facebook : la vie téléchargée", *OWNI*, 18 Novembre 2010 <http://owni.fr/2010/11/18/facebook-la-vie-telechargee/>

Foucault M. (1968), Sur l'archéologie des sciences. Réponse au Cercle d'épistémologie Dans *Dits et écrits*, tome 1, Paris, Gallimard.

Gallezot, G., et O. Le Deuff. (2009). « Chercheurs 2.0 ? » *Les Cahiers du numérique* 5(2), p. 15–31.

Ginzburg, C. (1989). Traces. Racines d'un pradigme indiciaire. Dans *Mythes, emblèmes, traces. Morphologie et histoire* (pp. 139-180). Paris, Flammarion.

Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm. Data-intensive scientific discovery.* Redmon (WA), Microsoft Research.

Jansen, B. J., Spink, A. (2006). How are we searching the World Wide Web ? A comparison of nine search engine transaction logs. *Information Processing and Management , 42*, pp. 248-263.

Latour, B. (1985) Les « vues » de l'esprit. Une introduction à l'anthropologie des sciences et des techniques. *Culture technique*, (14), p. 4-30.

Merzeau, L. (2009). Du signe à la trace : l'information sur mesure. *Hermès* (53), p. 23-29.

Rieder, B. (2010). Pratiques informationnelles et analyse des traces numériques : de la représentation à l'intervention. *Etudes de communication. Pratiques informationnelles : questions de modèles et de méthodes* (35), p. 91-103.

Salaün, J.-M. (2006). S'inspirer de Roger T. Pédauque. Dans R. T. Pédauque, *Le document à la lumière du numérique* (pp. 17-23). Caen, C&F éditions.

Serres, A. (2002). Quelle(s) problématique(s) de la trace ? *séminaire du CERCOR La question des traces et des corpus dans les recherches en Sciences de l'Information et de la Communication, 13 oct.*

Simonnot Brigitte, (2012) *L'accès à l'information en ligne : moteurs, dispositifs et médiations*. Paris, Hermès Lavoisier.

Sparrow B., Liu J., Wegner D.-M. (2011), "Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips", *Science*, 5 August 2011, Vol. 333 no. 6043, pp. 776-778 http://www.sciencemag.org/content/333/6043/776

Spivack N. (2009). "Welcome to the Stream – Next Phase of the Web", 8 mai 2009. <http://www.novaspivack.com/uncategorized/welcome-to-the-stream-next-phase-of-the-web>

Véronis J., (2010) "Google : le plus grand corpus linguistique de tous les temps", *Technologies du langage*, 16 décembre 2010 <http://blog.veronis.fr/2010/12/google-le-plus-grand-corpus.html>

Wesh M. (2007). "The machine is Us/ing us", 8 Mars 2007, <http://www.youtube.com/watch?v=NLlGopyXT_g>