Web, texte, conversation et redocumentarisation

Jean-Michel Salaün

EBSI – Université de Montréal – CP6128, succ Centre-ville –

Montréal, QC, Canada H3C 3J7

Abstract

Search engines mainly use linguistic and statistic tools and implicitly regard the Web like a large, global and permanent talk. The Web disturbs the traditional order of documents like a new media taking its place between old ones. A process of redocumentarization is currently taking place.
Usually even with new tools, linguists work within the old world of documents. However they have a special responsibility in building the new one.

Résumé

Les moteurs de recherche utilisent principalement des outils linguistiques et statistiques et considèrent implicitement la toile comme le vaste texte d'une conversation mondiale et ininterrompue. Le Web remet en cause l'ordre documentaire comme un nouveau média s'installant sans ménagement parmi les anciens. Un processus de redocumentarisation est en cours.
Les linguistes dans leurs travaux s'appuient généralement sur l'ordre documentaire ancien, même avec des outils nouveaux. Ils ont pourtant une responsabilité particulière pour définir le nouveau.

Mots-clés : Web, Texte, Text, Document, Redocumentarisation, Redocumentarization.

1. Introduction

Le texte ci-dessous s'appuie très largement sur le travail collectif réalisé sous la signature Roger T. Pédauque (2006) au sein du Réseau Thématique Pluridisciplinaire 33 sur le document du CNRS, maintenant en sommeil.

La navigation sur le Web des internautes est aujourd'hui structurée principalement par la dynamique des moteurs de recherche et les modes de lecture sur le Web sont très différents de ceux du monde du papier. On peut s'en persuader facilement par les résultats des enquêtes sur les pratiques. Plus que d'autres, celles concernant les populations académiques sont révélatrices, car elles touchent le cœur même de la bascule du monde lettré.

Ainsi, une enquête menée dans six pays (Australie, Canada, Indes, Singapour, Royaume-Uni et États-Unis) par *OCLC*[1] (2006) sur les étudiants des premiers cycles universitaires montre que :

- 89% d'entre eux démarrent une recherche d'information par un moteur de recherche.
- 93% sont satisfaits ou très satisfaits de leur utilisation générale des moteurs.
- Les moteurs correspondent mieux à leur style de vie que les bibliothèques physiques ou même numérique et la correspondance est «presque parfaite»

[1] Online Computer Library Center.

Une autre étude (UCL 2008), commandée par la *British Library*, montre que 60% des utilisateurs des revues électroniques scientifiques ne regardent pas plus de trois pages et la grande majorité d'entre eux n'y retournent jamais. Sans doute, un certain nombre télécharge les articles, mais il n'y a aucune preuve qu'ils les lisent. Les internautes passent une grande partie de leur temps à chercher leur chemin, bien plus qu'à regarder ce qu'ils ont trouvé. De plus, le temps moyen passé sur les sites de livres et revues électroniques est très court (respectivement 4 et 8 minutes). Les auteurs de l'étude remarquent (trad JMS) :

> *Il est clair que les internautes ne lisent pas en ligne dans le sens traditionnel du terme. Il y a en effet des signes manifestes que de nouvelles façons de «lire» apparaissent accompagnant les gains rapides de la «puissance de navigation» des internautes au travers des titres, du contenu des pages et des résumés. On pourrait presque dire qu'ils vont en ligne pour éviter de lire au sens traditionnel du terme.* (p.10)

On peut se demander si les internautes n'ont pas en fait engagé une conversation avec leur machine ou l'ensemble du système de représentation auquel elle donne accès.

2. Texte et conversation mondiale et ininterrompue

Les moteurs utilisent principalement des outils linguistiques et statistiques et considèrent, en effet, implicitement la toile comme le vaste texte d'une conversation mondiale et ininterrompue. Tous les genres, toutes les formes de publications ou d'échanges sont confondus dans un traitement uniforme qui marie une analyse de la langue et, de plus en plus, un calcul des flux. Ainsi, une requête sur un moteur peut ramener indifféremment un fragment de livre ou sa notice catalographique, un article de journal, un billet ou même un simple commentaire de blogue, une réponse archivée sur une liste, une rubrique d'encyclopédie, un argumentaire commercial, ou encore bien d'autres genres de textes. Sur le même moteur, le résultat est contextualisé, variant selon l'endroit et le moment où la requête est formulée.

Un exemple très simple suffit à l'illustrer. Une requête sur son nom faite sur le principal moteur amène toujours des réponses au statut documentaire hétéroclite différent selon qu'elle est formulée sur les sites canadien, français ou étatsunien de *Google*.

Google.ca	Google.fr	Google.com
• Mon blogue et un billet • Deux billets d'autres blogues • Un article de revue professionnelle • Deux articles de médias canadiens • Deux documents de présentation du programme de mon école • Une fiche de lecture de la bibliothèque de mon université sur l'article de la revue professionnelle.	• Mon blogue et un billet • Un article de revue professionnelle • Un autre de la revue du Cnrs • Deux fiches d'un libraire en ligne (qui, par ailleurs, a une publicité sur la page) • Trois notices biographiques • Une fiche d'un autre libraire en ligne	• Mon blogue et un billet • Deux billets d'autres blogues • Un article de revue professionnelle • Un autre de la revue du Cnrs • Deux fiches d'un libraire en ligne • Une notice d'annuaire en anglais • Un article de Webzine

Figure 1 : Résultat des dix premières réponses à la requête Jean-Michel Salaun le 15 février 2008

La requête des internautes, elle-même, est formulée le plus souvent en texte libre, sans hiérarchie, confirmant la posture conversationnelle entre l'internaute et un vaste robot piochant dans des textes. Le responsable de l'algorithmique des requêtes de *Google* a indiqué, dans une conférence récente (Manber 2007) que 20 à 25% des requêtes quotidiennes sur le moteur étaient nouvelles, soulignant par ce chiffre le caractère spontané et non réfléchi de l'usage de l'outil.

Un autre responsable de la firme (Varian, 2008) a précisé combien l'amélioration de l'efficacité des requêtes est dépendante de l'analyse du comportement de l'utilisateur (trad JMS) :

> *Au fil des années, Google a continué à investir dans de meilleures fonctionnalités de recherche. Nos experts de la recherche d'information ont ajouté plus de 200 nouveaux critères aux algorithmes qui déterminent la pertinence des sites Web pour la requête d'un utilisateur.*
>
> *Alors, d'où proviennent donc ces 200 critères supplémentaires ? Quelle est la prochaine étape de la recherche ? Et que nous devons faire pour trouver des informations encore plus pertinentes en ligne ?*
>
> *Nous ne cessons d'expérimenter notre algorithme, le réglons et le précisons sur une base hebdomadaire pour améliorer la pertinence et l'utilité des résultats pour nos utilisateurs.*
>
> *Mais pour trouver de nouvelles techniques de classement et évaluer si les utilisateurs trouvent leur bonheur, nous devons stocker et analyser les logs recherche.*

Mieux ou pire, le numérique permet aussi un mariage et une production à une échelle inédite entre les images, le son et l'écrit. Cet entrelacement a été préparé par la montée de

l'audiovisuel tout au long du XXème siècle, mais la convergence les trois supports de représentation explose littéralement sous nos yeux avec une transformation radicale des statuts de la photo, de la vidéo et de la musique, soutenue par les appareils numériques portables et multifonctionnels. Ces supports habituels du témoignage et de la distraction, dans l'ancien ordre documentaire, sont ainsi «démocratisés». Ils ont alors un usage encore peu socialement contrôlé. Les polémiques et les scandales sont légion. Il est inutile d'insister.

De plus en plus, ces usages passent par l'outil le plus conversationnel existant, le téléphone mobile, dont la rapidité d'appropriation par le grand public a dépassé, de très loin, celle de toutes les technologies qui l'ont précédé. On considérait à la fin de l'année 2006 qu'il y avait déjà 2,6 Mds de clients mobiles dans le monde, dont les 2/3 dans les pays en voie de développement (IDATE, 2007).

Mais si le Web favorise une transgression de l'ordre documentaire ordinaire, inversement il «documentarise» des expressions qui relevaient autrefois de l'intime et de l'éphémère en les enregistrant, les «traçant», les indexant. Le Web favorise conjointement deux mouvements autrefois opposés : le développement d'échanges spontanés (conversations) et leur fixation sur un support public, pérenne et documenté.

L'uniformisation entre conversation et document est encore plus flagrante, si on remarque que les frontières entre son propre espace documentaire privé (son «bureau»), la communication dans un groupe, une famille, une collectivité, une organisation, et l'espace public s'effacent de plus en plus. On écrit, on photographie, on publie, on partage, on réagit, tout cela «dans le ciel», c'est-à-dire potentiellement à la vue de tout le monde et on participe ainsi à une vaste conversation mondiale, ou on a l'illusion de le faire.

En réalité, on sait encore peu de choses sur le concret de ces pratiques et les premières études montrent qu'une minorité y est très active tandis que la majorité se contente de regarder (Prieur & alii, 2008). Mais la génération des internautes nés avec le numérique (Prensky, 2001) radicalise la rupture par une appropriation adolescente et un usage frénétique des jeux vidéos, des messageries instantanées, des téléphones mobiles et des «réseaux sociaux», comme leurs ainés se servaient de l'automobile ou du cinéma dans les années soixante, pour tourner la page de l'ancien monde de leurs parents.

3. Web-média

La façon la plus simple, et peut-être la plus rassurante, d'interpréter cette situation explosive pour nos représentations est de considérer que nous assistons à la naissance d'un média de masse nouveau. Ainsi l'ordre documentaire ancien serait écrasé par un média dont la naissance, et la brutale installation entre les modèles traditionnels de la bibliothèque et de la télévision, s'accompagnent à la fois de fulgurances et de confusions pour le sens commun. Cette interprétation est rassurante, car on peut supposer qu'une fois la poussière retombée, un nouvel équilibre sera trouvé entre les différents médias, induisant un ordre documentaire, sans doute différent et décalé, mais stable.

Pour appuyer cette thèse on peut remarquer que les résultats financiers de ceux qui ont su, corrélativement aux services qu'ils proposaient, construire le modèle d'affaires adéquat, ont explosé. Pour rester dans le monde académique, la firme *Reed-Elsevier* est devenue le premier éditeur mondial toute catégorie en terme de chiffre d'affaires (5,85 Mds d'Euros en 2006, 6,7 en 2007) (Livre-Hebdo, 2007), grâce notamment aux portails de revues scientifiques qu'elle a construits et aux licences d'accès qu'elle fait payer très cher aux bibliothèques universitaires.

À l'extérieur du monde académique, c'est évidemment *Google* qui rafle la mise et *the winner takes all* selon la formule traditionnelle de l'économique numérique. *Google* en 2007 a eu un chiffre d'affaires de 16,59 Mds de USD, pour un bénéfice de 4,2 Mds. Ce chiffre d'affaires est réalisé à 99% par la publicité, c'est-à-dire la captation de l'attention des internautes et sa revente à des annonceurs. 64 de ces 99% (proportion en augmentation année après année) proviennent des sites de la firme, grâce principalement au système *Adwords*, la vente aux enchères de mots-clés selon lesquels seront réparties ensuite les annonces en fonction du prix payé, le reste provient de l'activité de régie publicitaire pour d'autres sites. *Google* est entrée en 2007 dans le cercle des 10 plus grosses entreprises mondiales en termes de capitalisation boursière, performance remarquable pour une firme qui n'a que dix années et dont le secteur d'activité n'est pas plus ancien, même si le début de 2008 paraît être moins favorable.

Chaque média, bibliothèque, édition de livres, presse, radiotélévision s'est imposé en ébranlant l'ordre économique et la légitimité de ses prédécesseurs. Le Web ne fait pas exception, à ceci près que, cette fois, la rudesse du choc pourrait produire des dégâts plus lourds qu'auparavant. L'uniformisation pointée dans la partie précédente peut aussi, en effet, se comprendre comme le bouclage d'une histoire longue des médias, brusquement accélérée au siècle dernier.

Sans entrer dans une démonstration, nous pouvons illustrer cette thèse par un encadré schématisant les modèles économiques des médias.

1. **Édition**

On accroche l'information au support qui devient alors échangeable.

2. **Radio-télévision**

On capte l'attention du destinataire que l'on peut vendre à des émetteurs intéressés

3. **Bibliothèque**

On mutualise l'information entre les destinataires

1+2 : **Presse** 2+3 : **Web-média** ?

Figure 2 : Les trois modèles de médias et leur hybridation

Le Web, dans sa partie documentaire, emprunte au dernier né - la radio-télévision - bien de ses caractéristiques : un réseau, une vocation universelle dans son audience et son contenu, et surtout son économie marchande : la captation de l'attention de l'usager pour sa revente à des annonceurs intéressés. Mais il s'appuie aussi très largement sur des caractéristiques du plus ancien des médias - la bibliothèque – avec l'accès et la mise en réseau des collections, la distribution particulière des requêtes, entre conformisme et curiosité, rebaptisée aujourd'hui «longue traîne», et surtout le second versant de son économie : la mutualisation des objets documentaires.

Chaque média nouveau a été dénoncé par les gardiens de l'ordre ancien comme dévoyant les valeurs de la culture, du savoir ou de la démocratie. En délégitimant ce qui était légitime, en

publicisant ce qui n'aurait pas du l'être ou qui simplement ne l'était pas dans l'ordre documentaire précédent, le nouveau média perturbe la situation documentaire avant de trouver progressivement sa place dans le paysage général.

On retrouve pour le Web-média les mêmes polémiques qui ont accompagné la naissance de la presse populaire, celles de la radio et de la télévision. Il est un peu tôt pour savoir si celui-là s'assagira à terme, mais on peut déjà repérer facilement à la fois des avancées fortes qui bousculent une situation sclérosée par des hiérarchies dépassées ou simplement limitée par des outils trop rudimentaires et aussi quelques dérapages dangereux. Parmi de très nombreux autres exemples possibles et pour rester dans le domaine illustratif de notre introduction, l'évolution de la publication scientifique, jusqu'aux projets de cyberinfrastructures actuels (NSF, 2007) ou encore la place dominante prise par *Wikipédia* chez les étudiants (Whithe, 2007) en sont des illustrations.

4. Redocumentarisation

Mais cette première thèse, même si elle contient sans doute une part importante de l'explication, ne rend pas suffisamment compte de l'ampleur des bouleversements en cours. En effet, c'est la notion même de document qui se trouve aujourd'hui ébranlée, et avec elle les valeurs et les pratiques qui lui étaient attachées. Pour l'illustrer devant vous, le plus simple est peut-être de vous questionner sur vos propres pratiques de recherche.

Les chercheurs, qui se penchent sur les discours ou encore la structure de la langue grâce aux facilités de calcul ouvertes par l'informatique, ne s'interrogeaient pas beaucoup jusqu'ici, me semble-t-il, sur la légitimité des textes qu'ils étudiaient. Celle-ci était donnée par leur statut documentaire et ce dernier était le plus souvent garanti par la forme de l'objet et le traitement qu'il avait subi en amont, ce que l'on peut appeler sa «documentarisation».

Documentariser, c'est ni plus ni moins "traiter un document" comme le font, ou le faisaient, traditionnellement les professionnels de la documentation (bibliothécaires, archivistes, documentalistes) : le cataloguer, l'indexer, le résumer, le découper, éventuellement le renforcer, etc. On préfère "documentariser" à "documenter", qui renvoie plutôt à la création d'un ou de plusieurs documents pour expliquer un objet ou une action, mais dans nombre de cas les deux activités se recoupent. L'objectif de la documentarisation est d'optimiser l'usage du document en permettant un meilleur accès à son contenu et une meilleure mise en contexte. Par extension, on peut dire que documentariser, c'est donner un statut à un texte, le faire «document».

La documentarisation s'est appuyée sur un ordre documentaire issu de l'imprimé et, même s'il ne s'agit là que de sa partie la plus organisée, elle a été systématisée à partir du tournant du XXème siècle par la normalisation de règles de description et de classification, aussi bien du côté des bibliothécaires que de celui des archivistes ou des savoirs administratifs. Pendant tout le siècle dernier et encore aujourd'hui, l'effort s'est poursuivi et élargi avec l'évolution des techniques d'impression.

Ainsi les chercheurs linguistes ont réuni pour leurs besoins propres des corpus numériques, souvent d'ailleurs par l'opportunité du moment car ils étaient rares. Les corpus représentaient, de fait et implicitement, des collections organisées par une vie documentaire antérieure. Les collections, construites donc dans le monde de l'imprimé, ou analogues à ce dernier, induisaient une quasi-homologie entre texte et document, le premier étant inscrit de façon pérenne sur un support qui lui-même représentait l'objet documentaire.

L'évolution des formats numériques, avec, par exemple, la généralisation du XML qui sépare structure et texte ASCII, autorise des transpositions sur des supports de plus en plus hétérogènes et fragilise la solidité documentaire ancienne en proposant un autre découpage. Le développement des réseaux et l'explosion du Web, qui relient entre eux les fichiers et les textes, qui facilitent des modifications en continu transforment la nature documentaire de ces derniers. Les usages sociaux explosifs et les stratégies économiques lourdes conduisent à remettre en cause les classifications anciennes. Nous ne sommes plus donc dans cette situation confortable où texte et document étaient presque synonymes.

Les conséquences sont sans doute importantes pour les chercheurs qui doivent se poser sur les matériaux qu'ils analysent des questions inédites, au sens propre, car ceux-ci ont peut-être perdu le statut implicite qu'ils avaient auparavant. L'objet ayant changé de nature, on peut s'interroger pour savoir s'il est pertinent de l'analyser comme s'il ne s'était rien passé. Vous avez sûrement sur cette question des réponses plus appropriées que je ne saurais en avoir, mais je crois utile de se la poser.

Plus globalement, les conséquences sont aussi très lourdes sur l'ensemble de notre relation au savoir enregistré et peut-être même sur l'ensemble des relations sociales si l'on se réfère au rôle général de régulation joué par l'ordre documentaire.

C'est pourquoi, il est essentiel de tenter de comprendre la dimension documentaire des mouvements en cours. Globalement, nous assistons à un mouvement massif et désordonné de «redocumentarisation» dont il est difficile de prévoir l'issue tant elle est, chaque fois que l'on croit l'apercevoir, contredite par d'intenses échanges et brassages qui déstabilisent les tentatives d'explication. Les appels et les efforts lancés par Tim Berners-Lee pour un Web sémantique, puis pour une science du Web, peuvent être compris dans cette perspective. Mais bien d'autres acteurs et bien d'autres logiques se croisent.

Le parallèle entre la situation actuelle et celle qui prévalait au début du siècle dernier au moment de la systématisation de la documentarisation est tout à fait frappant. Je reproduis ci-dessous un tableau déjà paru dans un article précédent (Salaün, 2007) sur le phénomène, qui pointe quelques unes de ses caractéristiques sans prétendre en faire le tour.

	Documentarisation	**Redocumentarisation**
Dates	Tournant XIXe-XXe	Tournant XXe-XXIe
Quelques figures	M. Dewey, P. Otlet, O. Lafontaine, W Carnegie	T. Berners-Lee, T. Nelson, B. Gates, S. Brin
Quelques techniques	Classification, Indexation, Langages documentaires, Thésaurus..	Protocoles Web (Html, Url) Web 2.0, Web sémantique Ontologies..
Quelques réalisations	Réseau mondial de bibliothèques	Google, Wikipédia
Les modernités	L'esprit scientifique, la raison-logique, l'État-nation, les votes, l'industrie, l'auteur ..	Le savoir limité, la raison-statistique, l'individu, les opinions, les services, la réflexivité ..
Quelques objets documentaires concernés	Les revues, les règlements, les contrats, les brevets, les œuvres, les médias et l'imprimerie.	Les pré-publications, les formulaires, les sources ouvertes, les wikis, les blogues et le web

Figure 3 : Les deux bascules documentaires

Si l'on reprend le domaine scientifique qui nous a servi d'introduction, on peut penser que les mouvements et travaux sur ce qu'il est convenu d'appeler l'e-science et l'interopérabilité des données relèvent de ce processus (Lynch, 2007).

5. Conclusion

Dans cette présentation, j'ai avancé trois idées complémentaires pour éclairer l'explosion de l'ordre documentaire ancien qui se déroule sous nos yeux et la place des linguistes :

1. Les acteurs dominant du Web le considèrent comme le vaste texte d'une conversation mondiale ininterrompue.
2. Le Web remet en cause l'ordre documentaire comme un nouveau média s'installant sans ménagement parmi les anciens.
3. Une tentative de redocumentarisation est en cours.

Dans ce processus, les linguistes ont une responsabilité particulière, car les outils qu'ils développent avec les informaticiens sont au cœur du mouvement. Peut-être le moment est-il venu pour eux de s'interroger sérieusement sur l'importance de la notion de document pour structurer nos mémoires collectives.

Références

Google. (2007). *Google Announces Fourth Quarter And Fiscal Year 2006 Results*. Communiqué janvier 2007. http://investor.google.com/releases/2006Q4.html

Google. (2008). *Google Announces Fourth Quarter And Fiscal Year 2007 Results*. Communiqué janvier 2008. http://investor.google.com/releases/2007Q4.html

IDATE. (2007). *Digiworld2007 les enjeux du monde numérique*. *http://idate.fr/pages/index.php?title=&idrbis=&rubr=&nummenu=&rubrique=digiworld&idl=6&idr=15&idp=24&download=ok&id=45&rub=digiworld_telech&nom=DW2007FR.pdf*

Livre-Hebdo. 2007. *Le classement 2007 de l'édition mondiale*. Brochure. http://www.livreshebdo.com/cache/upload/pdf

Lynch C. (2007). The Shape of the Scientific Article in The Developing Cyberinfrastructure. *CTWatch Quaterly*. 3(3). http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/

Manber U. (2007). Meaningful Search Queries, Conférence *Supernova2007,* 21 juin 2007. http://itc.conversationsnetwork.org/shows/detail3360.html

NSF (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation Cyberinfrastructure Council. Mars 2007. http://www.nsf.gov/od/oci/CI_Vision_March07.pdf

OCLC (2006). *College Students' Perceptions of the Libraries and Information Resources*. Rapport. Dublin, OH. http://www.oclc.org/reports/pdfs/studentperceptions.pdf

Pédauque R. T. (2006). *Le document à la lumière du numérique*. C&F éditions.

Prensky M. (2001). Digital Natives, Digital Immigrants, *On the Horizon*, 9(5).

http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf

Prieur C. & alii *The Stength of Weak cooperation: A Case Study on Flickr*, papier de recherche, 16 février 2008. http://arxiv.org/ftp/arxiv/papers/0802/0802.2317.pdf

Salaün J.-M.. (2007). La redocumentarisation, un défi pour les sciences de l'information. *Études de communication*. 30. http://hdl.handle.net/1866/1724

UCL. (2008). Information behaviour of the researcher of the future. Rapport commandé par la British Library et JISC. http://www.bl.uk/news/pdf/googlegen.pdf

Varian H. (2008). Why data matters, *The Official Google blog*, 3/04/2008. http://googleblog.blogspot.com/2008/03/why-data-matters.html

Withe D. (2007). Results of the "Online Tool Survey" undertaken by the JISC funded SPIRE project. http://tallblog.conted.ox.ac.uk/wp-content/uploads/2007/03/survey-summary.pdf