Du document à la donnée et retour

La fourmilière ou les Lumières

Jean-Michel Salaün

Version de travail à publier dans *Le document numérique à l'heure du web de données : séminaire INRIA, 1er au 5 octobre 2012, Carnac*. Paris, ADBS Éditions, 2012

Chercheur en sciences de l'information, Jean-Michel Salaün est professeur à l'École normale supérieure de Lyon où il est en charge du premier master francophone sur l'architecture de l'information. De 2005 à 2010, il était directeur de l'École de bibliothéconomie et des sciences de l'information (EBSI) de Montréal. Il a été auparavant professeur à l'École nationale supérieure des sciences de l'information et des bibliothèques (ENSSIB) en France, où il a pris diverses responsabilités, dont l'animation du réseau du CNRS RTP-Document (Roger T. Pédauque). jeanmichel.salaun@ens-lyon.fr

Même si le mot est ancien, la référence à la notion de document est récente dans l'Histoire, sans doute en résonance avec l'organisation de la société industrielle, sa régulation et ses valeurs[1]. Au tournant du millénaire, le web s'est appuyé sur un renversement du circuit documentaire, jusqu'à, dans le web de données, un court-circuitage radical. S'agit-il de l'effacement d'une notion périmée au profit d'une autre ou d'un simple décalage ? Le succès du web accompagne des transformations sociales et économiques profondes. Quelles seront alors, sur nos régimes de vérité, de preuve et de transmission, les conséquences de l'éventuelle obsolescence de la notion de document ?

1 Bref rappel sur la notion de document

Ayant déjà eu l'occasion de m'exprimer sur la notion de document, je ne présenterai ici qu'un très bref rappel, renvoyant le lecteur qui voudrait éclaircissements et approfondissements à d'autres publications sur le sujet [16] [25] [26].

Depuis le Moyen Âge, un document est un artefact qui sert à transmettre des connaissances et à témoigner comme preuve. Mais la notion de document n'est devenue vraiment courante qu'à partir du XIX^{e} siècle et évidemment au XX^{e} pour répondre aux besoins de la science moderne, du commerce, de l'industrie, et à la transformation de la bureaucratie des États. À partir de la fin du XIX^{e} siècle un effort de documentarisation systématique et international, c'est-à-dire de récolte, catalogage et indexation de l'ensemble des documents publiés à l'échelle de la planète, a été entrepris avec le développement des plus importantes bibliothèques universitaires, la rationalisation des bibliothèques nationales ou encore l'ouverture de grands centres de documentation. C'est à cette époque que les premières interrogations sur la notion ont été posées par le Belge Paul Otlet avec son projet de documentariser l'ensemble des publications du monde [20].

Avec la montée de l'informatique, puis l'avènement du web, ce qu'il est convenu maintenant d'appeler le numérique, la réflexion sur le document a été renouvelée pour accompagner le processus devenu indispensable de redocumentarisation, c'est-à-dire de révision radicale de la documentarisation précédente. Le document changeant de support et subissant les manipulations d'autres outils, il a fallu en mieux comprendre la structure et la fonction pour l'intégrer dans son nouvel environnement. Dans ce renouvellement de la réflexion, un groupe de chercheurs interdisciplinaires au sein du CNRS a proposé d'analyser le document selon trois dimensions, trois facettes, à la fois autonomes et articulées : la forme, le texte, et la fonction [21], que j'ai résumées de façon mnémotechnique par trois participes passés : le vu, le lu et le su [25].

Dans cette proposition, un document n'existe que si ces trois dimensions sont mises en cohérence par un contrat de lecture qui relie les producteurs ou les responsables du document et ses usagers, ses lecteurs. Un passeport, un livre, un testament, un journal, une facture, un article scientifique ne deviennent documents pour un couple de producteur et lecteur que si, dans un contexte donné, ils peuvent être repérés comme tels (vus), on peut en interpréter le contenu (lus) et enfin ils ont une fonction de médiation (sus). Vous saurez repérer un passeport, le voir d'un coup d'œil parmi divers papiers imprimés sur une table ; si vous le prenez et le lisez, vous pourrez indiquer le nom de la personne à qui il appartient et son pays d'origine ; et vous savez que cette personne en a besoin pour voyager, prouver son identité et sa nationalité. Cela n'est possible que parce qu'il existe un contrat de lecture que vous avez intégré au point qu'il fait maintenant partie de vos réflexes mentaux. Celui-ci vous permet d'appréhender cet objet comme un passeport, un document ayant une forme particulière, un contenu précis et une fonction connue.

[1] Cette contribution doit beaucoup aux échanges avec mes camarades pédauquiens : Bruno Bachimont (UTC-INA), Valérie Beaudouin (Télécom ParisTech), Jean Charlet (Inserm), Michael Eberlé-Sinatra (Université de Montréal), Jean-Philippe Magué (ENS de Lyon), Yves Marcoux (Université de Montréal), Alain Mille (Université Claude Bernard - Lyon 1), et tout spécialement Benoît Habert (ENS de Lyon) dont les remarques ont été déterminantes.

Chaque dimension joue son rôle avec des importances variables selon les époques. La seconde dimension est devenue primordiale au moment de la popularisation de la notion de document. Le texte, en effet, joue un rôle central depuis au moins le XVIII^{e} siècle. L'imprimerie qui le fixe à l'identique sur un support en de nombreux exemplaires n'est pas pour rien dans cette priorité. Le texte doit rester stable. Comme l'a montré Elizabeth Eisenstein, cette caractéristique a favorisé le développement des sciences et de l'esprit critique à l'époque moderne [11]. On lui attache, par exemple, un droit moral inaliénable dans le droit d'auteur ou encore on en atteste la validité par une signature, certifiée par un tiers assermenté, comme un notaire, par exemple, pour les documents ayant des conséquences juridiques. Ainsi on peut dire que cette seconde dimension (le lu, celle du texte) constitue, pour l'époque moderne, le cœur de la notion de document.

Bien des questions qui se jouent aujourd'hui viennent de l'ébranlement de cette stabilité par les développements récents du web, le web 2.0 d'abord, puis le web de données. La thèse que je voudrais suggérer ici est que le web peut se comprendre comme une tentative d'adapter la notion de document à la modernité tardive, pour reprendre l'expression d'Hartmut Rosa [24], c'est-à-dire la modernité qui caractérise nos sociétés occidentales contemporaines. Autrement dit, le web comme l'imprimerie à son époque est venu à son heure pour participer aux transformations de notre rapport aux connaissances et les accélérer, pour le meilleur et pour le pire.

2 Le web des documents

À son origine, le web dans les années 90 est conçu comme un service documentaire. Si les innovations proposées par Tim Berners-Lee tranchent par rapport aux médias existants, bibliothèques comprises, elles ne bouleversent pas vraiment leur matière première : le document dont l'intégrité du texte n'est pas mise en cause.

Le premier web, dit « web des documents » s'appuie sur le langage HTML, lui-même issu de SGML, dont une des caractéristiques est de séparer la structure du document de son contenu, faisant ressortir par là-même une de ses propriétés : l'autonomie relative de la forme par rapport au texte, qui sera confirmée et systématisée par le succès du langage XML. Autant la première, la forme, conserve une certaine plasticité dans les limites des normes visuelles communément admises pour un genre donné (un même document peut être reproduit sous différentes représentations), autant le second, le texte, ne souffre pas de transformation (modifier le texte dénature le document moderne, comme nous l'avons rappelé). Pour reprendre le vocabulaire de la partie précédente : le numérique souligne que la partie du contrat de lecture relevant de la première dimension a une certaine souplesse, tout en confirmant que, au contraire, la contractualisation sur la deuxième dimension fixe une rigidité : l'intangibilité du texte.

La seconde innovation du premier web concerne les liens entre les documents à partir du texte lui-même. Reprenant l'intuition de Vannevar Bush sur les associations d'idées [8], telle que l'avait déclinée un peu plus tard Eugene Garfield à partir des citations dans les articles scientifiques [13], le web relie les documents ou des points (ancres) à l'intérieur de documents. Cette innovation perfectionne le document comme nœud de réseau sans pour autant remettre en cause le texte lui-même. On passe en un clic d'un point d'un texte à un autre sans modifier ni le premier, ni le second.

La troisième innovation ne figurait pas dans le projet initial mais a découlé très vite de la dynamique même du réseau documentaire : il s'agit des moteurs de recherche. Tim Berners-Lee, l'inventeur du web, indique : « Dès le développement du Web, ses détracteurs ont souligné qu'il ne pourrait jamais être une bibliothèque bien organisée, que, sans base de données centrale et sans structure arborescente, on ne pourrait jamais être sûr de tout trouver. Ils avaient raison. Mais la puissance d'expression du système a mis à la disposition du public des quantités importantes d'informations et les moteurs de recherche (qui auraient paru tout à fait irréalisables il y a dix ans) permettent de trouver des ressources [6]. »

De plus, le principal moteur de recherche, dans sa version gagnante, s'appuie non seulement sur un traitement linguistique des textes mais aussi sur leur classement issu du nombre de liens menant à un document, c'est-à-dire sur une application généralisée du principe d'Eugene Garfield : le fameux *PageRank* de Sergey Brin et Larry Page, fondateurs de Google [7]. Le moteur de Google est une formidable machine à lire des documents, mais elle ne remet pas en cause l'intégrité des textes qui sont au contraire, par leur stabilité même, sa matière première.

Ainsi le premier web, baptisé par Tim Berners-Lee « web des documents », s'inscrit dans la continuité du document moderne. Il en tire un meilleur profit grâce aux protocoles mis en place, rendus possibles par les performances toujours accrues des transistors sur le silicium et du signal numérique, qui autorisent une souplesse encore plus grande que celle du papier imprimé ou du signal analogique. Les principales caractéristiques du document moderne ne sont pas pour autant remises en cause.

3 Web 2.0 et web de données

Les étapes suivantes du développement du web ont des conséquences plus radicales sur le document. Je ne dirai que quelques mots du web 2.0 pour insister sur le web de données, sujet principal de ce livre.

L'expression web 2.0 a été inventée et popularisée non par les promoteurs du web, ceux qui inventent les normes et protocoles au sein du consortium W3C, mais plutôt par ceux qui se sont emparés de l'outil pour inventer de nouveaux services, en premier lieu par l'éditeur Tim O'Reilly [18]. Du point de vue documentaire, il s'agit bien d'une nouvelle étape pour le web. En insistant sur le partage, la conversation, la réactivité, c'est-à-dire en privilégiant la troisième dimension du document, sa fonction sociale, ces acteurs ont ébranlé le pilier du document moderne : l'intangibilité du texte. Dans le web 2.0, le droit d'auteur est contesté par la pratique de l'échange, de la copie, du mixage. La notion même d'auteur est relativisée par la pluralité des interventions sur les textes. Et surtout leur stabilité est mise à mal par les corrections et ajustements

possibles et courants. On retrouve même dans les néodocuments du web 2.0, les billets de blogues, les wikis, les pages personnelles alimentés par les amis, etc., nombre de caractéristiques des *documens* du Moyen Âge d'avant l'imprimerie : la leçon commentée, la copie assortie des corrections du copiste, la glose, etc. Ce qui a fait dire à un chercheur que nous sortions de la « parenthèse Gutenberg [22] ».

En résumé le développement des services de partage sur le web, menés indépendamment de la feuille de route des acteurs techniques du W3C, débouche sur une révision du document moderne, comme en témoignent nombre de débats souvent difficiles entre les promoteurs d'un échange libéré et les représentants du système documentaire ancien.

La troisième étape du web, qu'il est convenu aujourd'hui d'appeler le « web de données » après l'avoir nommé « web sémantique » [12], est une remise en cause plus radicale encore de la notion de document. En effet, il ne s'agit plus de partager des documents, quitte à en pervertir la stabilité comme dans le web 2.0, mais bien de reconstruire automatiquement et à la demande des documents à partir de données partagées. Dès lors, la notion même de texte est questionnée.

Mais, paradoxalement et de façon implicite, la logique du processus reprend un cheminement proche des pratiques et traditions documentaires construites à la fin du XIX^{e} siècle, celles de la documentarisation, comme le montrent les présentations de son promoteur. Pour atteindre son objectif d'un web de données, Tim Berners-Lee pose en préalable la mise à disposition libre des données. La diapositive reproduite en figure 1 est extraite d'une de ses conférences [5], sans doute la plus diffusée, où il défend avec conviction cette proposition.

Cette image illustre le processus de constitution de la base de données DBpedia à partir de l'extraction de données des rubriques de l'encyclopédie Wikipédia. Les données factuelles essentielles de chaque rubrique sont réunies dans une *infobox*, placée en haut et à droite de la page pour informer rapidement le lecteur et aussi permettre aux robots de traiter facilement ces données qui sont structurées. Le projet DBpedia consiste à réunir ces données dans une base de données centrale et à les rendre interopérables.

L'image devrait éveiller quelques souvenirs aux bibliothécaires. Wikipédia, en effet, réalise une opération qui s'apparente au catalogage, rédigeant une notice avec des champs structurés (*infobox*), puis un catalogue (DBpedia). L'homologie est encore plus évidente et surprenante si l'on se souvient que Paul Otlet, un des premiers théoriciens du document, avait lui aussi proposé la réalisation d'une encyclopédie, dans son rêve de cataloguer tous les documents du monde [20, p. 41]. L'encyclopédie et la classification décimale universelle constituaient les deux instruments principaux de la documentarisation pour retrouver les connaissances contenues dans les documents récoltés.

Néanmoins, il existe bien une différence radicale entre le projet de Paul Otlet et celui de Tim Berners-Lee. La documentarisation s'est déplacée. Pour le premier, l'enjeu est de récolter les documents pour les cataloguer, et l'encyclopédie est un aboutissement, en réalité une utopie qui ne sera jamais vraiment opérationnelle. Pour le promoteur du web, l'objectif de la récolte des documents est atteint par la dynamique même du réseau. Les documents sont déjà en ligne, repérables par les moteurs de recherche. Mieux, l'encyclopédie est aussi déjà là et s'enrichit elle aussi dans une dynamique continue : Wikipédia est devenu une figure emblématique du web 2.0, au succès spectaculaire et mondial.

L'enjeu est alors différent, inversé même. Il s'agit de reconstruire des documents en se servant en priorité des données récoltées et formatées notamment dans l'encyclopédie en ligne et aussi dans d'autres bases coopératives encyclopédiques comme Freebase. DBpedia est considérée par les chercheurs comme « un noyau pour un web de données ouvertes [2] ». Tim O'Reilly souligne pour sa part que Freebase constitue le pont entre l'intelligence collective issue de la base du web 2.0 et le monde plus structuré du web sémantique [19].

Les métadonnées ne servent plus à retrouver un document : « libérées », rendues interopérables et traitables par les logiciels du web de données, elles se détachent de leur document d'origine pour se combiner et produire de nouveaux documents. Dans le web de données, l'unité de base n'est plus le document comme précédemment, mais la donnée. On ne relie plus des documents entre eux, mais des données entre elles. Le langage dominant n'est plus celui qui permet de rédiger un texte et de le relier à d'autres (HTML), mais celui (RDF) qui permet de décrire et relier entre elles les données – c'est-à-dire les entités du monde – par un triplet[2] qui constitue en quelque sorte l'unité textuelle de base. Le lien entre les données autorise la constitution de documents à la volée selon la navigation de l'internaute et pour éclairer son action.

À cette première base de données ouvertes et normées peuvent alors se relier nombre d'autres bases de toutes natures, libérées elles aussi et pouvant être mieux décryptées grâce à l'ossature initiale de DBpédia. Il se constitue ainsi un graphe de données reliées entre elles. La figure 2 présente la version simplifiée et partielle du graphe ainsi constitué, au centre duquel se trouve la base DBpedia[3].

Il n'est pas indifférent que l'encyclopédie Wikipédia ait été un outil privilégié pour construire l'armature de ce nouvel ensemble. Tout se passe comme si les wikipédiens avaient catalogué les entrées de l'encyclopédie en ligne et que la mise en réseau de cette classification sinon universelle, comme celle qu'ambitionnait Paul Otlet, du moins largement partagée pouvait autoriser la construction d'un nouveau monde documentaire. Dans le développement du mouvement de l'*Open data*, lancé par Tim Berners-Lee et ses collègues, ce sont les données publiques qui sont logiquement de plus en plus sollicitées. Déjà nombre de collectivités, gouvernementales ou locales, ont ouvert l'accès aux données qu'elles récoltent et produisent. La combinaison entre la vocation de bien commun central qui est celle de Wikipédia et cette galaxie de bases publiques reconfigure petit à petit un espace public d'un genre nouveau où les connaissances ordinaires et utiles sont calculées et proposées à la demande et rendues accessibles pour le citoyen branché grâce à un contrat de lecture renouvelé.

---

[2] Entité-propriété-valeur (par exemple : F. Hollande-né en-1954) ou entité1-propriété-entité2 (F. Hollande-président-République française).

[3] Voir la figure complète dans le chapitre 3, figure 6, page XX.

4 Le Knowledge Graph

Il est trop tôt pour imaginer l'avenir et mesurer les conséquences du projet du web de données sur notre rapport au document, même si on peut penser qu'elles seront importantes puisque la remise en cause de la stabilité du texte est ici radicale. Mais il est déjà symptomatique de retrouver parmi les tout premiers à développer une application grand public issue directement de la logique du web de données le même acteur qui avait permis au premier web, celui des documents, d'acquérir les qualités d'un service de bibliothèque : Google.

En effet, Google a annoncé tout récemment l'intégration d'une dimension sémantique dans son moteur de recherche, baptisée « Knowledge Graph[4] ». Pour cela, il a indiqué qu'il avait développé un algorithme pour puiser des informations dans des bases de données comme Freebase ou DBpedia. Le principe du service est de proposer, parallèlement aux réponses traditionnelles aux requêtes (liste de liens vers des documents pertinents récupérés par le *PageRank*), des informations construites à la volée donnant des éléments de contexte et, le cas échéant, la réponse elle-même. On trouvera en figures 3 une illustration du résultat en comparant la requête « François Hollande » sur Google.fr [fig. 3a], qui n'a pas encore intégré le nouveau service, et sur Google.com [fig. 3b].

La page de réponse est profondément modifiée par rapport à sa version antérieure puisque la partie droite est constituée d'un encadré issu du traitement des données disponibles et proposant un portrait résumé. La partie gauche n'est pas non plus sans leçons pour nous puisqu'on y trouve immédiatement un lien sur la rubrique de Wikipédia correspondante, comme pour la plupart des requêtes sur le moteur. Google affiche aussi implicitement cette connivence avec l'encyclopédie collaborative dans le communiqué présentant le nouveau service puisque, dans les trois exemples proposés, la première référence à apparaître est toujours Wikipédia.

Cette connivence n'est pas nouvelle. On peut même prétendre que, sans Google, Wikipédia n'aurait pas eu le même succès et que, sans Wikipédia, Google n'aurait pas eu la même saveur. Mieux, la tentative de Google de construire sa propre encyclopédie a tourné court, comme si le succès de la coopération dans Wikipédia supposait l'affichage du désintéressement. Les deux services sont alors complémentaires, fondant un écosystème au sens fort du terme basé sur l'économie de l'attention. Mais aujourd'hui l'omniprésence de Wikipédia, aussi bien dans la construction collective des rubriques que dans les réponses traditionnelles aux requêtes des moteurs que, enfin, dans la construction automatique de documents par DBpédia selon le principe du web de données, souligne son caractère de plus en plus central dans le processus documentaire de la modernité tardive, du moins pour les interrogations courantes.

Pour Google et le Kwowledge Graph, l'utilisation des données du web du même nom se combine avec celles que la firme récolte sur les comportements des internautes, pour à la fois personnaliser les réponses aux requêtes et affiner sa vente de mots clés aux annonceurs. Comme le suggère Olivier Andrieu sur son blog spécialisé dans le référencement [1], il semble que l'impact sur le nombre de requêtes ait été immédiat. Il ajoute : « Cela est logique dans le sens où le "Knowledge Graph" propose de nombreux liens concernant l'objet de la requête et de "l'entité nommée" détectée. D'ailleurs, cela pourrait clairement être à l'avantage de Google : l'internaute tape une requête sur la page d'accueil du moteur, obtient les résultats du Knowledge Graph, reclique sur les liens de recherche proposés, etc. Bref, autant de possibilités d'afficher des Adwords pour Google et autant de clics publicitaires potentiels ! » Ajoutons que le processus tend à réduire l'impact des stratégies de référencement des sites au profit de l'achat de mots-clés et, corollairement, à faire monter dans les réponses les rubriques de Wikipédia.

5 Les données, une histoire à écrire

Le web et le numérique en général ont été l'occasion d'insister sur le développement des données, par le projet de web de données que nous venons d'évoquer, mais aussi par la mise en mémoire des données numériques de toutes sortes, phénomène aujourd'hui connu sous le vocable de *big data*[5]. Néanmoins la récolte et l'utilisation des données n'ont pas attendu le numérique pour se développer dans les sociétés modernes. Depuis les recensements de populations, les tableaux économiques, l'avènement des bourses et places financières, le développement des observations dans diverses disciplines scientifiques, l'analyse des marchés puis le marketing, les sondages divers jusqu'à l'utilisation des données pour piloter les machines dans la cybernétique, les données se sont multipliées et leur traitement est devenu courant dans tous les rouages du fonctionnement des sociétés contemporaines.

L'histoire des données reste encore à écrire, mais il existe des travaux importants du côté de l'histoire sociale de la statistique, notamment ceux d'Alain Desrosières [10] qui montre combien celle-ci est corrélée à la consolidation des États modernes et plus généralement à la représentation d'une société par elle-même et à l'organisation de son espace public.

Alain Desrosières insiste notamment sur l'articulation entre deux niveaux, celui de la collecte et celui du traitement des données, chacun étant issu d'une logique particulière et aucun n'étant neutre dans la production des connaissances : « Dans son architecture actuelle, la statistique se présente comme la combinaison de deux types d'outillages distincts, dont les trajectoires historiques n'ont convergé et conduit à une construction robuste que vers le milieu du XX^{e} siècle. Le premier est politico-administratif : peu à peu se sont mis en place, depuis le XVIII^{e} siècle, des systèmes d'enregistrement, de codage, de tabulation et de publications de "statistiques" au sens de description chiffrée de divers aspects du monde social. Le second est cognitif, et implique la mise en forme de schèmes scientifiques (moyenne, dispersion, corrélation, échantillonnage

[4] Communiqué de Google : http://googleblog.blogspot.fr/2012/05/introducing-knowledge-graph-things-not.html.

[5] Pour une première approche, voir la présentation d'Hubert Guillaud [15].

probabiliste) destinés à résumer, notamment par des outils mathématiques, une diversité supposée non maîtrisable [10, p. 398]. »

Ainsi ce n'est pas seulement d'une histoire contemporaine des données dont nous aurions besoin, mais d'une compréhension croisée entre l'histoire de leur récolte et celle de leur traitement, c'est-à-dire des algorithmes[6] pour éclairer l'analyse du nouvel outillage qui se met en place sous nos yeux. Comme pour les statistiques, ni l'une, ni l'autre ne sont neutres pour notre rapport aux connaissances et tout particulièrement aux documents.

Il n'est pas indifférent de remarquer que la mise en place progressive des premiers outils statistiques et la montée de la notion de document sont contemporaines, tout en résultant de logiques distinctes. Comme si les institutions des sociétés modernes avaient besoin conjointement, pour s'imposer, d'outils permettant de les décrire « objectivement » et d'artefacts simples et souples permettant la transmission des connaissances et la représentation de la preuve. Une photographie reposant sur la collecte et le calcul d'un côté, un texte de l'autre. Citons une nouvelle fois Alain Desrosières : « La constitution d'un espace rendant possible le débat contradictoire sur les options de la cité suppose l'existence d'un minimum d'éléments de référence communs aux divers acteurs : langage pour mettre en forme les choses, pour dire les fins et les moyens de l'action, pour en discuter les résultats. Ce langage ne préexiste pas au débat : il est négocié, stabilisé, inscrit, puis déformé et défait peu à peu, au fil des interactions propres à un espace et une période historique donnés [10, p. 406]. »

De ce point de vue, le web de données constitue bien une nouvelle étape. Les données ne sont plus simplement des éléments factuels permettant de conforter ou d'infirmer la rhétorique des textes, mais bien les unités textuelles de base. Le texte lui-même est calculé, pourrait-on dire, ou, pour reprendre la proposition de Bruno Bachimont [3] à la suite de Jack Goody [14], nous passerions d'une « raison graphique », celle de l'écriture, du texte, à une « raison computationnelle », celle du calcul par les algorithmes.

6 Transparence et accélération

Thomas Bern, de son côté, montre qu'à la fin du XVI[e] siècle se tient en France à propos du recensement une polémique dont les arguments sur la transparence prennent un relief étonnant aujourd'hui [4]. Il oppose ainsi le principe de publicité, celui de la discussion publique des Lumières conduisant à la loi, à celui de transparence, celui de la réalité et de l'objectivité du calcul induisant la norme. Nous pourrions dire, en suivant cet auteur, que l'artefact « document » est plutôt en phase avec le premier principe tandis que l'artefact « donnée », comme son nom l'indique, relèverait plutôt du second. Or, poursuit-il, la transparence conduit à un gouvernement « inoffensif » où les décisions politiques ne sont plus discutées et sont diluées dans le social.

Bien des thématiques ou des postures actuelles autour du web font un écho singulier à cette affirmation : la neutralité de point de vue revendiquée dans la rédaction des rubriques de Wikipédia, la « sagesse des foules » pour le filtrage des informations, l'exhortation à libérer les données sans recul critique quant aux conditions de leur récolte ou aux conséquences de leur croisement, la neutralité affirmée des algorithmes des principales firmes assurant la gestion documentaire du web jusqu'au monde sans conflit du *don't be evil* de Google, des « amis » de Facebook ou des *followers* de Twitter, etc.

Une autre conséquence de la dynamique du web sur les relations au document devrait nous amener à réfléchir. À la suite de Paul Virilio, Hartmut Rosa considère que la « modernité tardive » est caractérisée par une accélération aussi bien de notre vie quotidienne que des changements sociaux ou encore des techniques, pour lui accélération rime avec aliénation [24]. Le développement du web, et tout particulièrement celui de la combinaison du web de données et du traitement de nos traces, participe très largement à cette accélération où toutes les réponses à nos questions sont fournies à peine posées et parfois avant même d'être posées[7]. Certains prétendent que « 90 % de l'ensemble des données du monde ont été créées ces deux dernières années[8] ». Or cette accélération a pour conséquence de réduire notre capacité à penser ou à réfléchir puisque la lecture se transforme pour devenir superficielle, si l'on en croit Nicolas Carr [9], ou que le débat perd souvent de son enjeu, réduit à des affrontements rapides, et se concluant bien souvent à peine ouvert sans convaincre.

Transparence et accélération dessinent un monde un peu effrayant, bien loin du bazar caractéristique de l'économie du logiciel libre [23], ou encore de l'abeille dont le butinage autorise la pollinisation des fleurs dans l'économie du web [17]. La métaphore pertinente serait plutôt la fourmilière où nos échanges et nos rapports documentaires auraient pour vocation unique de faciliter notre vie quotidienne, elle-même partie prenante d'un grand tout qui nous dépasserait et dont la finalité nous échapperait. Le lecteur aura compris que je n'évoque ce tableau que pour mieux le conjurer.

Inversement, on peut considérer, à l'instar de Michel Serres [27], que les développements actuels du web nous libèrent de contingences documentaires en délestant notre mémoire de connaissances courantes aujourd'hui directement accessibles sur le réseau et nous donnent du temps pour réfléchir et prendre des décisions plus pertinentes parce que mieux informées. Le web serait alors l'occasion d'un renouveau des Lumières.

Entre le cauchemar de la fourmilière et le rêve des Lumières, il est probable que l'Histoire nous montrera une troisième voie. Il me semble que l'approfondissement d'une théorie du document, désormais élargie à celle des données, pourrait nous aider

---

[6] On peut saluer sur ce point les premières analyses critiques de Tarleton Gillepsie et de Dominique Cardon.

[7] Voir sur ce sujet le nouveau service de Google, Google Now, présenté sur Rue 89 (2 juillet 2012), d'où je tire cette citation d'un des dirigeants de la firme, Éric Schmidt : « En réalité, je pense que les gens ne veulent pas que Google réponde à leurs questions. Ils veulent que Google leur dise ce qu'ils doivent faire. » http://www.rue89.com/2012/07/02/sur-android-google-veut-repondre-aux-questions-quon-ne-lui-pose-pas-233529.

[8] Représentant d'IBM cité par Hubert Guillaud [15].

à mieux la dégager et peut-être à éviter quelques pièges. L'Histoire, et tout particulièrement celle de l'Europe du XX^{e} siècle, nous a appris qu'il n'y avait pas plus de déterminisme documentaire qu'il n'y a de déterminisme technique. Les performances documentaires ont pu être mises au service des totalitarismes nazis, fascistes ou staliniens comme à celui de l'émancipation des peuples. Nous sommes face à des défis différents, mais les transformations du document ne sont pas un simple symptôme secondaire accompagnant le succès du web. Il s'agit de l'ébranlement de l'artefact qui nous sert à transmettre des connaissances et à témoigner comme preuve, ou de « la trace qui nous permet d'interpréter un événement passé à partir d'un contrat de lecture [25] ». Cela n'est pas anodin.

Références

[1] Olivier ANDRIEU. « Google : le "Knowledge Graph" génère plus de requêtes... et de publicités ». *Abondance Actualité*, 4 juin 2012. http://www.abondance.com/actualites/20120604-11526-google-le-knowledge-graph-genere-plus-de-requetes%E2%80%A6-et-de-publicites.html

[2] Sören AUER *et al*. « DBpedia: A Nucleus for a Web of Open Data ». *In* : K. ABERER *et al*. (dir.). Actes du colloque *ISWC/ASWC 2007*. Berlin : Springer-Verlag, 2007. P. 722–735. http://www.springerlink.com/content/rm32474088w54378/fulltext.pdf?MUD=MP

[3] Bruno BACHIMONT, *Arts et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle*. Habilitation à diriger des recherches : Université de technologie de Compiègne, 12 janvier 2004. 281 p. http://www.utc.fr/~bachimon/Livresettheses_attachments/HabilitationBB.pdf

[4] Thomas BERN, « Transparence et inoffensivité du gouvernement statistique ». *Raison-Publique.fr*, 11 juillet 2011. http://www.raison-publique.fr/article447.html

[5] Tim BERNERS-LEE. *On the next Web*. Conférence TED, février 2009. http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

[6] Tim BERNERS-LEE, James HENDLER et Ora LASSILA. « The semantic Web, A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities ». *Scientific American Magazine*, 17 mai 2001. http://www.scientificamerican.com/article.cfm?id=the-semantic-web. [Traduit par Élisabeth LACOMBE et Jo LINK-PEZET : http://www.urfist.cict.fr/archive/lettres/lettre28/lettre28-22.html]

[7] Sergey BRIN et Lawrence PAGE. « The anatomy of a large-scale hypertextual web search engine », *In* : *Seventh International World-Wide Web Conference*. Brisbane, Australia, 14-18 april 1998. http://infolab.stanford.edu/~backrub/google.html

[8] Vannevar BUSH. « As we may think ». *The Atlantic*, juillet 1945. http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/1

[9] Nicholas CARR, *The Shallows. What the internet is doing to our brains*, New-York : Norton, 2010. 276 p.

[10] Alain DESROSIERES, *La politique des grands nombres : histoire de la raison statistique*. 2^{e} éd. Paris : La Découverte, 2000. 462 p.

[11] Elizabeth L. EISENSTEIN. *La Révolution de l'imprimé dans l'Europe des premiers temps modernes*. Paris : La Découverte, 1991. 360 p.

[12] Fabien GANDON. « Technologies et architecture du web de données ». *Documentaliste - Sciences de l'information*, 2011, vol. 48, n 4, p. 27-30.

[13] Eugene GARFIELD. « Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas ». *Science*, juillet 1955, vol. 122, n° 3159, p. 108-11. http://garfield.library.upenn.edu/papers/science1955.pdf

[14] Jack GOODY. *La raison graphique : la domestication de la pensée sauvage*. Paris : Éditions de Minuit, 1979. 272 p.

[15] Hubert GUILLAUD. « Vers un nouveau monde de données ». *Internet Actu*, 1er juin 2012. http://www.internetactu.net/2012/06/01/vers-un-nouveau-monde-de-donnees

[16] Niels Windfeld LUND et Roswitha SKARE. « Document theory ». In : *Encyclopedia of Library and Information Sciences.* 3^{e} éd. New York : Taylor and Francis, 2010. P. 1632-1639

[17] Yann MOULIER-BOUTANG. *L'abeille et l'économiste*. Paris : Carnets Nord, 2010. 254 p.

[18] Tim O'REILLY. "What Is Web 2.0. Design Patterns and Business Models for the Next". *O'Reilly Radar*, 30 septembre 2005. http://radar.oreilly.com/pub/a/web2/archive/what-is-web-20.html

[19] Tim O'REILLY. "Freebase Will Prove Addictive". *O'Reilly Radar*, 8 mars 2007. http://radar.oreilly.com/2007/03/freebase-will-prove-addictive.html

[20] Paul OTLET. *Traité de documentation : le livre sur le livre. Théorie et pratique.* Bruxelles : Van Keerberghen, 1934. [Reprod. en fac-sim. : Liège : Centre de lecture publique de la Communauté française de Belgique ; Bruxelles : Ed. Mundaneum-Palais mondial, 1989]

[21] Roger T. PEDAUQUE. *Le Document à la lumière du numérique. Forme, texte, médium : comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité.* Caen : C&F éditions, 2006. 226 p.

[22] Tom PETTITT. « Before the Gutenberg Parenthesis: Elisabethan-American Compatibilities ». *Media in Transition 5: Creativity, Ownership and Collaboration in the Digital Age. Pleinière 1: "Folk Cultures and Digital Cultures"*. MIT, non daté. http://web.mit.edu/comm-forum/mit5/papers/pettitt_plenary_gutenberg.pdf

[23] Eric S. RAYMOND. *La cathédrale et le bazar*. 11 août 1998. http://www.linux-france.org/article/these/cathedrale-bazar/cathedrale-bazar_monoblock.html

[24] Hartmut ROSA. *Accélération : une critique sociale du temps*. Paris : La Découverte, 2010. 480 p. (Théorie critique)

[25] Jean-Michel SALAÜN. *Vu, lu, su. Les architectes de l'information face à l'oligopole du Web*. Paris : La Découverte, 2012. 151 p. (Cahiers libres)

[26] Jean-Michel SALAÜN. « Pourquoi le document importe ». *Les E-Dossiers de l'audiovisuel*, juin 2012. http://www.ina-sup.com/node/2832

[27] Michel SERRES. *Petite poucette*. Paris : Éditions du Pommier. 89 p.

Légendes des figures

Figure 1 – La constitution de DBpedia

Figure 2 – Les bases de données ouvertes et connectées

Figure 3a : Réponse de Google.fr à la requête « François Hollande »

Figure 3b : Réponse du Knowledge Graph (Google.com) à la requête « François Hollande »