Le catalogue des bibliothèques et ses données à l'heure du web

Raphaëlle Lapôtre

▶ **To cite this version:**

Raphaëlle Lapôtre. Le catalogue des bibliothèques et ses données à l'heure du web. 2016. halshs-01331753

HAL Id: halshs-01331753
https://shs.hal.science/halshs-01331753v1

Preprint submitted on 29 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Le catalogue et ses données à l'heure du web

Raphaëlle Lapôtre, juin 2016

"Le web a pour effet immédiat de créer une économie de l'abondance d'information (...)[1]" écrivait Emmanuelle Bermès en 2013. Force est de constater aujourd'hui que se poser la question du positionnement du catalogue sur le web revient toujours nécessairement à se poser celle de son rôle dans l'économie de l'attention telle que définie depuis le début des années 2000 par les géants du web : Internet est en effet présenté aujourd'hui comme le parangon de ce versant de l'économie qui voit en l'attention publique un bien rare et précieux dont il est nécessaire de gérer l'usage. Cependant, si l'on définit l'interface utilisateur du web comme étant à la fois une représentation du monde et une mesure de l'attention à ce même monde, comment théoriser l'usage qu'en font ses principaux acteurs ? A cet égard, il peut être démontré qu'au moins deux visions de l'économie de l'attention s'opposent et se complètent : reste à savoir dans quel segment de ces théories affrontées se positionnent les catalogues des bibliothèques. Il est par ailleurs étonnant de constater que ces visions économiques peuvent être décrites selon les catégories proposées par Michel Foucault en 1966[2] pour décrire la pensée économique des XVII^{e} et XVIII^{e} siècle : dans ce contexte, l'attention ne serait qu'un élément s'insérant dans un système plus vaste qui pourrait être celui de l'ordre économique de la représentation. Cette proposition que nous faisons est peut-être confortée par le fait que les données sur le web et les données du catalogue s'analysent tout aussi bien en tant que monnaie dans un cadre économique qu'en tant que langage dans un cadre sémantique.

La donnée-monnaie

Tout d'abord, il nous faut revenir sur les propos d'Emmanuel Kessous[3] qui écrivait, en analysant les propos de Michael H. Goldhaber que "l'économie de l'attention [peut être] définie comme une ère nouvelle succédant à l'économie féodale et à l'économie industrielle fondée sur le marché et la circulation monétaire", ajoutant plus loin que dans cette économie "les concepts usuels de l'économie perdent toute signification"[4]. S'il est vrai que les concepts de production et de travail ne s'appliquent guère à l'économie de l'attention, en revanche, peut-on passer sous silence le rôle prédominant des données et de leur organisation sur l'écran ? À cet égard, peut-être faudrait-il parler, plutôt que d'attention au monde, d'attention aux signes qui tout à la fois désignent les objets du monde ainsi que la valeur de ces objets dans l'ordre général des richesses : la mise en scène des données sur l'écran exprime la

---

[1] BERMES, Emmanuelle. *Le web sémantique en bibliothèque*. Cercle de la librairie, 2013. p. à compléter.

[2] FOUCAULT, Michel. *Les mots et les choses : une archéologie des sciences humaines*. Paris : Gallimard, 1966.

[3] KESSOUS, Emmanuel, *L'attention au monde : sociologie des données personnelles à l'ère numérique*. Paris : Armand Colin, 2012.

[4] *Ibidem* p. 163-164.

valeur attribuée par l'acteur du web, qu'il soit moteur de recherche, simple site internet ou blog, aux objets désignés par ces signes. Ainsi, la position du signe sur la page, entre le haut et le bas, la gauche et la droite, manifestent un "rang" - on pense bien évidemment au PageRank des créateurs de Google - à savoir un prix dont l'attention peut être vue comme l'étalon, c'est-à-dire le bien précieux et rare à l'aune duquel cette valeur est mesurée. C'est ainsi que les données à l'ère du web rejoignent la définition attribuée au XVII^{e} siècle par le mercantilisme à la monnaie, à savoir un ensemble de signes universels dont la fonction est à la fois de représenter les richesses, et par là de se substituer temporairement à elles, ainsi que de mesurer leur valeur, le métal précieux ne servant qu'à justifier d'une certaine manière la fonction mesurante du signe, au même titre que l'intérêt subjectif et supposé des internautes ne servirait en quelque sorte qu'à justifier la disposition des signes sur la page dans l'ordre économique du web. Ce n'est d'ailleurs pas la première fois que l'élément monétaire est utilisé pour décrire le fonctionnement de certains acteurs du web. Citons à ce titre la description que fait Dominique Cardon du Science Citation Index, "cette représentation" qui, "tout en préservant un lien référentiel avec le monde qu'elle enregistre, (...) invente aussi un cadre cognitif bien particulier dont on [peut] dégager cinq propriétés épistémiques qui appareilleront le Page Rank"[5]. En effet le SCI, en considérant toute référence, sortie de son contexte, comme une citation représentant en quelque sorte le texte auquel elle fait référence "invite à considérer toutes les citations comme égales. Cette transformation, ajoute Dominique Cardon, contribue donc à unifier le sens de la citation pour en faire une sorte de "monnaie de l'activité scientifique", à la fois standardisée, décontextualisée, univoque et égale[6]." Ainsi, du SCI au PageRank, de la citation scientifique au lien hypertexte, de la "monnaie de l'activité scientifique" à la monnaie de l'activité documentaire dans laquelle circule l'internaute, il n'est qu'un pas, rapidement franchi par Larry Page dans les années 2000.

Si l'on compare les procédés de deux acteurs majeurs et complémentaires de la recherche d'information sur le web aujourd'hui que sont Google et Wikipédia, on peut observer que pour la même requête ambiguë que pourrait être par exemple "Charles VIII", la manière de lever l'ambiguïté entre Charles VIII de France et Charles VIII de Suède n'est pas la même d'un acteur à l'autre. En effet, là où Google s'appuie sur le parcours précédent de l'usager sur le web, sur le contenu de ses mails ainsi que sur le parcours des personnes dont il aura relevé une proximité avec lui sur les réseaux, et lui propose une liste de résultats par laquelle il a classé l'entité estimée la plus susceptible de correspondre à sa recherche, Wikipedia se révèle quant à lui agnostique et redirige l'utilisateur sur une page d'homonymie dans laquelle il sera invité à choisir l'entité correspondant à sa recherche parmi toutes les possibilités offertes dans l'ordre du savoir défini par l'encyclopédie en ligne. Or, pourrait-on dire qu'une méthode se révélerait plus efficace qu'une autre pour retrouver l'information recherchée ? Certes, on pourrait nous reprocher de comparer l'incomparable, en mesurant les performances d'un moteur de recherche généraliste à celle d'une encyclopédie en ligne, pour autant les questions posées à ces deux outils peuvent être les mêmes. Il est intéressant de noter que d'un point de vue technologique, rien ne paraît empêcher le célèbre moteur de recherche de la firme Mountain View de proposer des facettes classificatoires pour affiner les premier résultats d'une requête, de même qu'aucune contrainte technologique ne viendrait apparemment entraver la possibilité pour Wikipédia de rediriger

[5] CARDON, Dominique. "Dans l'esprit du PageRank." *Réseaux* no. 177. 2013. p 68.
[6] *Ibidem*. p. 69.

l'utilisateur sur une de ses pages en tenant seulement compte de son parcours sur son domaine : il semblerait pour autant que cela ne soit pas inscrit dans leur philosophie respective. Se confirment ainsi les propos d'Andrew Feenberg, pour qui "la conception technologique n'est pas déterminée par un critère général qui serait celui de l'efficacité, mais par un processus social qui distingue les alternatives techniques en fonction d'une variété de critères spécifiques à chaque cas", ajoutant plus loin que "des définitions rivales reflètent des visions opposées de la société moderne qui se réalisent dans des choix techniques différents.[7]" Que l'on parle d'algorithmes classificatoires ou de présentation graphique de la page, ces deux procédés technologiques, piliers du fonctionnement du web aujourd'hui, impliquent la transcription de conceptions du monde variées dans un langage symbolique, puisque d'un côté les algorithmes peuvent être vus comme la pensée "littéralement faite mécanique, en se fixant dans les signes que nous encodons de manière précise et logique[8]" et de l'autre, "la présentation de l'information peut être comprise et lue comme des expressions culturellement encodées du savoir, avec leur propre présuppositions épistémologiques et leur lignage historique[9]." En clair, lorsque nous avons affaire au web, nous avons affaire à du langage, et plus exactement, puisque l'algorithme précède la présentation visuelle, nous avons affaire à un langage qui se re-présente, un théâtre des signes dont la philosophie est à rechercher dans l'âge de la représentation, telle que défini par Michel Foucault en 1966 dans *Les Mots et les Choses*.

Reste à savoir quels peuvent être ces présupposés épistémologiques encodés par les technologies du web, et puisque des moteurs de recherche aux encyclopédies en ligne, nous parcourons le domaine de l'économie de l'attention, il nous semble que c'est dans l'ordre économique de la représentation qu'il nous est nécessaire de porter notre regard. Ainsi, analysant l'algorithme kNN utilisé par Google mais également par la plupart des sites web commerciaux faisant l'usage d'algorithmes de recommandation, Thomas Neal situe la philosophie économique de Google du côté de l'utilitarisme : "les interfaces web modernes permettent désormais la définition et la satisfaction de "situations problématiques" socialement contextualisées, de manière à pouvoir agir en tant que médiatrices d'une information transformatrice. Le point focal en est la production perpétuelle et collective de sens, et le cadre théorique est une correspondance utilitaro-économique entre le sujet et l'objet. Plus simplement, elles sont basées sur la théorie du choix rationnel[10]." Or

---

[7] FEENBERG, Andrew. *Questioning technology*. London: Routledge. 1999. p. 83-84.

[8] "The efficiency for human beings is to be found where thinking can literally be made mechanical, by fixing the signs we encode into computer in logically precise ways." THOMAS, Neal. "Algorithmic subjectivity and the need to be in-formed". In LATZKO-TOTH Guillaume, MILLERAND, Florence. *TEM 2012 : Proceedings of the Technology & Emerging Media Track – Annual Conference of the Canadian Communication Association (Waterloo, May 30 - June 1, 2012)*, http://www.tem.fl.ulaval.ca/www/wpcontent/PDF/Waterloo_2012/THOMAS-TEM2012.pdf, p. 3.

[9] "My basic aim is to create a critical framework within which the forms that are generally used for the presentation of information can be understood and read as culturally coded expressions of knowledge with their own epistemological assumptions and historical lineage." DRUCKER, Johanna. Graphesis: Visual knowledge production and representation. Poetess Archive Journal. 2010. Vol. 2, n°1, pp. 1–50. Consulté le 6 août 2014. Disponible à l'adresse Web http://www.johannadrucker.com/pdf/graphesis_2011.pdf. p. 4.

[10] "Contemporary network interfaces like Google rely on the collective posing and satisfaction of ongoing, socially contextualized 'problem situations', so that they can act as an intermediary for transformative information. The focus is on perpetual, collective sense-making, and the theoretical framework is a utilitarian-economic correspondence between subject and object. More simply, it is based on rational choice theory." NEAL, *Ibidem*. p. 4.

l'utilitarisme est ce qui, pour Michel Foucault, place la valeur d'un bien du côté de sa réception, c'est à dire de la mesure du besoin dont les hommes, placés en situation d'échange, en ont. Plus la nécessité d'un bien est grande, plus le bien a de valeur par rapport à d'autres biens. C'est ainsi que l'algorithme kNN ne se préoccupe pas de classer les biens eux-mêmes mais de classer les usagers qui en ont parcouru la représentation sur le web. De là naît la recommandation, qui mesure la valeur du signe à sa place dans le parcours de consultation des usagers : ceux qui ont consulté x ont également consulté y. À l'opposé de cette théorie utilitariste, mais toujours selon le même segment théorique, Michel Foucault situe la physiocratie, qui place la valeur non dans sa réception, mais dans le bien lui-même, dans le donné, ou plus exactement, dans le surplus de donné laissé après une opération de soustraction progressive par rapport à la consommation de ce bien : la valeur serait dans ce qui reste après consommation, elle « n'apparaît que là où des biens ont disparu [11]». Ainsi, du surplus de donné, nous passerions, dans l'ordre du web et de la représentation, au surplus de la donnée, c'est à dire à sa capacité, par abstraction et non plus par soustraction, à signifier un ensemble de biens. Ainsi, dans un catalogue FRBRisé, tant de documents existant d'Hamlet exemplarisent tant d'éditions qui sont la déclinaison de toutes ces expressions possibles d'Hamlet qui elles-mêmes peuvent être regroupées en une seule et même œuvre Hamlet. Pour le faire vite, dans l'ordre économique dessiné par les FRBR, la valeur du point de vue de la gestion de l'attention est dans l'œuvre et sa capacité par abstraction à signifier un ensemble plus ou moins grands de documents, l'œuvre n'étant autre que le dénominateur commun de signification formé grâce à la comparaison successive de toutes ses manifestations physiques.

À ces deux visions théoriques de l'économie de l'attention correspondraient des algorithmes les exprimant, avec d'une part le PageRank et kNN du côté des acteurs du web utilisant la recommandation pour signifier la valeur, d'autre part, du point de vue du monde de la documentation, le work-set algorithm dont s'est servi l'OCLC pour faire des essais de FRBRisation du catalogue WorldCat[12]. Quant au domaine de la visualisation de l'information, ces théories opposées s'expriment également dans deux types de présentation : du côté de l'utilitarisme et de la réception, la visualisation la plus fréquente est la liste pure et simple de résultats, chaque signe pouvant se retrouver d'un jour à l'autre en tête ou en queue, valorisé ou non dans l'ordre économique de la représentation. Du côté de la physiocratie et de la donnée-signe on retrouve la présentation en facettes, présente jusque sur les grands sites d'e-commerce que sont Amazon et e-bay et dont le principe est de déplier les abstractions successives jusqu'à parvenir à l'item recherché. Les deux principes traduits par ces algorithmes et ces présentations visuelles sont deux manières d'envisager la valeur dans l'ordre économique du web, l'une analyse la valeur du côté du besoin, de la carence d'information, c'est-à-dire du côté de celui qui est dans la recherche d'information, l'autre l'analyse du point de vue de la surabondance d'information (les multiples exemplaires d'une même œuvre) et désigne la valeur dans la transformation de cette surabondance en un symbole synthétique. Le fait que nous opposions des moteurs de recherches généralistes, pour lesquels la définition de l'objet informationnel est plutôt lâche (le texte libre de la barre de recherche) à des sites souvent spécialisés (le catalogue des bibliothèques ne s'intéresse

[11] FOUCAULT*, Ibidem*, p. 207.

[12] HICKEY, Thomas B., TOVES, Jenny. "FRBR Work-Set Algorithm. Version 2.0. Dublin, OH : OCLC Online Computer Library Center, Inc. (Research Division). Published online at : http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf, 2009.

qu'au « document ») montre bien deux conceptions opposées de l'accès à l'information : d'un côté la valeur attentionnelle est envisagée dans le regard porté sur elle et se soucie peu de la nature du bien observé. Les données véritablement traitées seront donc des données personnelles. De l'autre, cette même valeur est considérée du point de vue du découpage transformationnel qui fera de la profusion d'items une seule et même œuvre en passant par des manifestations et des expressions : les données traitées seront alors des métadonnées, à savoir des données qui ont pour fonction principale la désignation d'un objet. Il s'agira donc dans un cas d'attribuer au bien un regard valorisant ou non, dans l'autre, de découper la valeur dans la profusion des biens informationnels disponibles à l'attention : « dans le système des échanges, écrit Michel Foucault, dans le jeu qui permet à chaque part de richesse de signifier les autres ou d'être signifiée par elles, la valeur est à la fois verbe et nom, pouvoir de lier et principe d'analyse, attribution et découpe. »[13] En face des liens hypertextes et de la circulation du regard sur ces liens, nous aurions donc la structuration des données et leur pouvoir de désigner un nombre plus ou moins grand d'objets informationnels selon l'échelon où elles se trouvent dans l'ordre défini par leur modèle : en face du verbe donc, nous aurions le nom.

La prose du web

Si l'on reprend les deux conceptions économiques évoquées précédemment, à savoir ce que nous appellerions dans un premier temps "utilitarisme" et "physiocratie", et que nous examinons les différents artefacts technologiques qui leur correspondent - dans un cas un algorithme s'appuyant sur le lien hypertexte, dans l'autre des facettes classificatoires permettant de découper la représentation -, on s'aperçoit que ces deux modes de pensée sont la transcription de deux métaphores récurrentes dans le domaine de l'organisation des connaissances que sont le réseau et l'arbre. C'est ainsi que dans son ouvrage intitulé *De l'arbre au labyrinthe*, Umberto Eco retrace la généalogie de ces métaphores, rattachant la tradition de l'arbre à la pensée médiévale et celle du réseau ouvert aux Encyclopédistes, pour les analyser ensuite dans leur déclinaison informatique contemporaine que sont les ontologies d'une part et les liens hypertextes d'autre part. Or, au moment où Eco décrit l'oeuvre du philosophe et homme d'Église John Wilkins intitulée *An Essay towards a real character and a philosophical language*, il est fascinant d'observer qu'il décrit un système dont les principes ont été énoncés au XVII^{e} siècle et qui pour autant est très proche de ce que nous appelons aujourd'hui le web sémantique :

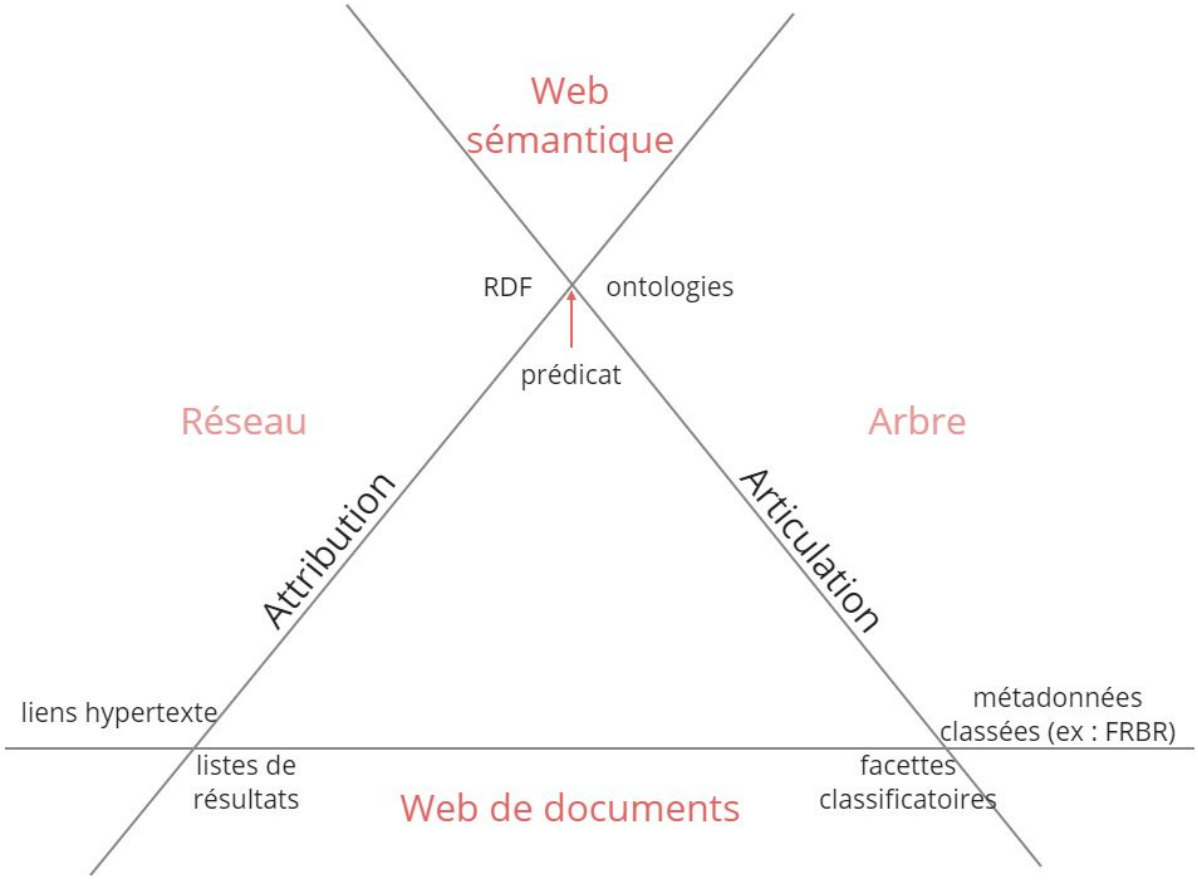
"(...) Le système wilkinsien, du fait justement de son impureté, pourrait donner lieu à une autre lecture, non plus comme un dictionnaire mais comme hypertexte, au sens actuel du terme En effet, si un hypertexte lie chaque noeud ou élément de son propre répertoire, à travers une multitude de renvois internes à de multiples noeuds, on peut alors concevoir un hypertexte sur les animaux qui fasse s'insérer chien dans une classification générale des mammifères, dans un arbre de taxa contenant également le chat, le boeuf et le loup. Mais si, dans cet arbre, on pointe sur chien (au sens informatique actuel : on clique sur), on sera alors renvoyé à un répertoire d'informations concernant les propriétés du chien, ses habitudes. En sélectionnant un autre ordre de connexions, on accèdera à une revue des différents rôles du chien à différentes époques historiques, ou à une énumération des images du chien dans l'histoire de l'art[14]".

[13] FOUCAULT, *Ibidem*. p. 215.
[14] ECO, Umberto. *De l'arbre au labyrinthe*. Paris : Grasset, 2010. p. 65.

Difficile, ici, de ne pas reconnaître non seulement le système de l'interopérabilité par les liens mais également la définition et classification des objets au sein de référentiels et ontologies qui caractérisent le web de données. En d'autres termes, nous pourrions dire que le projet du web sémantique s'inscrit de plein pied dans les projets de langue universelle qui ont caractérisé la pensée du XVII^{e} siècle, hypothèse pouvant être étayée par le fait qu'au-delà de la volonté de trouver un système de classification du réel, le projet de Wilkins était bien celui d'une langue qui permettrait une communication non ambiguë entre universitaires et philosophes. Ainsi, de même que pour le web sémantique, le prédicat est à la fois chargé de lier un sujet à un attribut, mais aussi de porter un jugement de classification sur son attribut (le prédicat renvoie à des propriétés définies au préalable dans une ontologie), de même le verbe "être" était-il pour les grammairiens de Port-Royal, en tant qu'essence même du langage et de la proposition, chargé d'attribuer un objet à un sujet tout en articulant les deux objets liés au sein d'une classification allant du général au particulier, du nom commun au nom propre.[15]

C'est ainsi que l'on retrouve unis au sein même du web sémantique les deux principes d'attribution et d'articulation qui caractérisaient les deux segments opposés de l'économie de l'attention, dont la manifestation technologique était d'une part un algorithme reposant sur les liens hypertextes et d'autre part une présentation de l'information sous forme de facettes classificatoires. D'une certaine manière, nous pourrions donc caractériser la logique représentationnelle du web sous la forme d'un triangle au sein duquel les deux segments de l'attribution et de l'articulation, opposés dans l'ordre économique du web, finiraient par se rejoindre dans l'ordre sémantique du web. On aurait ainsi la figure ci-dessous :

[15] "La généralité du nom est aussi nécessaire aux parties du discours que la désignation de l'être à la forme de la proposition.
Cette généralité peut être acquise de deux manières. Ou bien par une articulation horizontale, groupant les individus qui ont entre eux certaines identités, séparant ceux qui sont différents ; elle forme alors une généralisation successive des groupes de plus en plus larges (et de moins en moins nombreux) ; elle peut aussi les subdiviser presque à l'infini par des distinctions nouvelles et rejoindre ainsi le nom propre dont elle est partie ; tout l'ordre des coordinations et des subordinations se trouve recouvert par le langage et chacun de ces points y figure avec son nom : de l'individu à l'espèce, puis de celle-ci au genre et à la classe, le langage s'articule exactement sur le domaine des généralités croissantes ; cette fonction taxinomique, ce sont les substantifs qui la manifestent dans le langage : on dit un animal, un quadrupède, un chien, un barbet." FOUCAULT, *Ibidem*. p. 112-113.



Nous pourrions donc décliner le web et sa représentation du point de vue de trois croisements : une première intersection fondamentale, au sein du web sémantique, des principes d'attribution et d'articulation pourrait être vue au sein même du prédicat, qui porte en lui-même la faculté de lier deux entités et de définir ce lien du point de vue d'une ontologie. Puis, au croisement du web de documents et du principe d'attribution nous aurions le lien hypertexte qui déterminera l'affichage de la liste des résultats sur la page web tandis qu'au croisement opposé du triangle, le regroupement des données du général au particulier contribuera à une organisation de la page web reposant sur un classement défini au préalable par une ontologie.

Il paraît intéressant de noter que parmi les visualisations de données les plus fréquentes pour représenter le web sémantique figurent en bonne place le graphe hiérarchique et le graphe non-hiérarchique : souvent combinés, ces deux types de représentation matérialisent en quelque sorte les deux métaphores sous-jacentes du web que sont l'arbre et le labyrinthe. D'une certaine manière, nous sommes de nouveau en face d'un langage, le web sémantique, qui se représente lui-même en reproduisant dans la visualisation ses propres principes fondateurs. Ainsi, de même que le RDF figure par le biais d'un prédicat un lien entre deux entités, et par la position de ces entités autour du prédicat la direction de ce lien, de même le principe fondateur du graphe est de faire figurer deux points reliés ensemble par un trait dont l'extrémité fléchée pourra indiquer la direction du sujet vers l'attribut (cf figure n°1 ci-dessous).

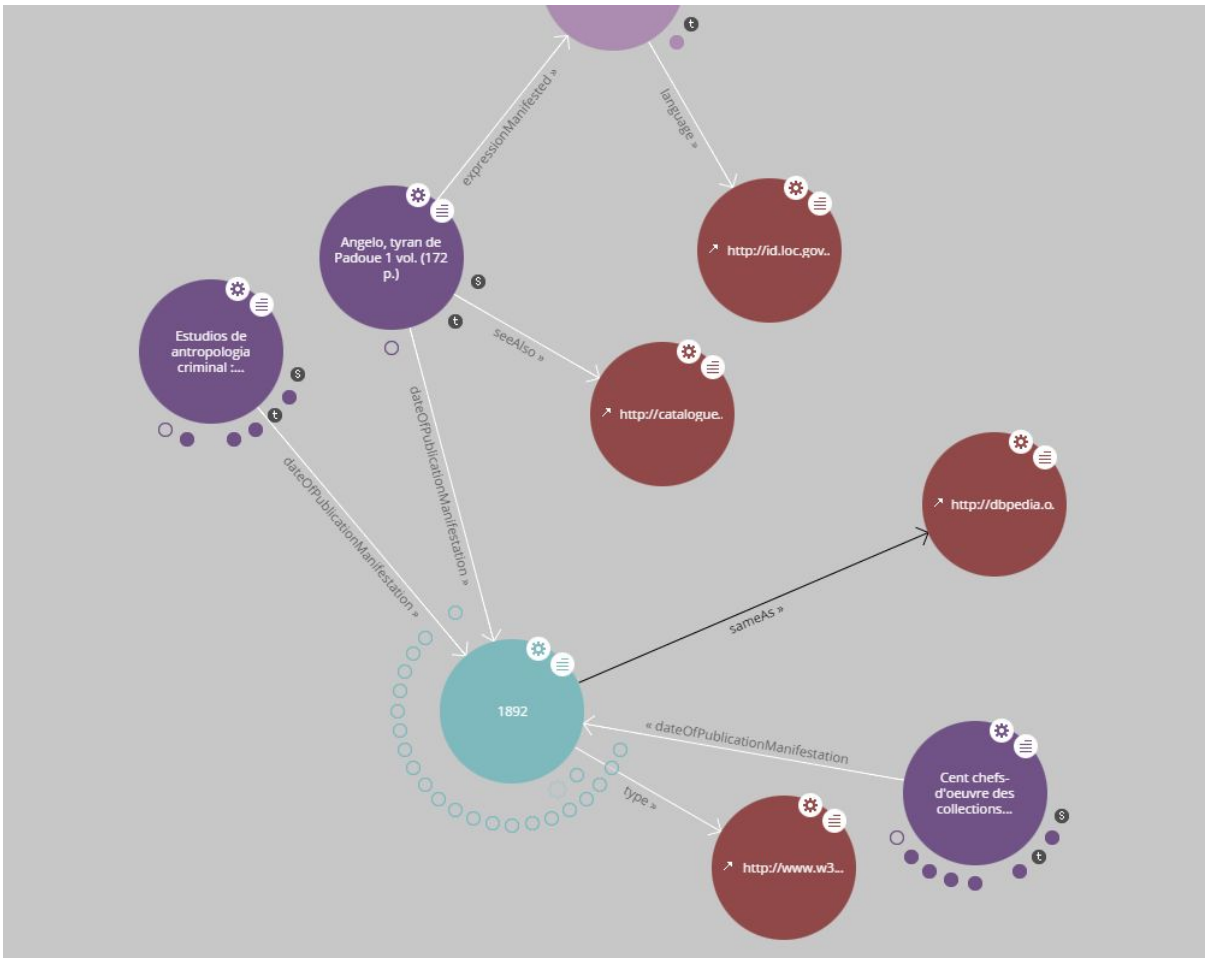


Figure n°1 : Visualisation d'un graphe RDF dans l'outil de visualisation en ligne lodlive

Ensuite, à l'image d'une ontologie dont le rôle est de définir entre quels types d'entités il est possible ou non d'établir tel ou tel type de lien, une visualisation arborescente ne pourrait plus fonctionner en tant qu'arborescence s'il arrivait qu'un des liens qu'elle figure reliait deux entités figurant sur deux branches opposées de l'arbre. Dans la figure n°2 ci-dessous par exemple, la création d'un nœud correspondant à une autobiographie qui pourrait être relié à la fois au nœud de documents "by the author" et à celui "about the author" serait perturbante : il ne serait plus alors possible de replier par un clic ce nœud additionnel dans l'un ou l'autre des nœuds supérieurs.

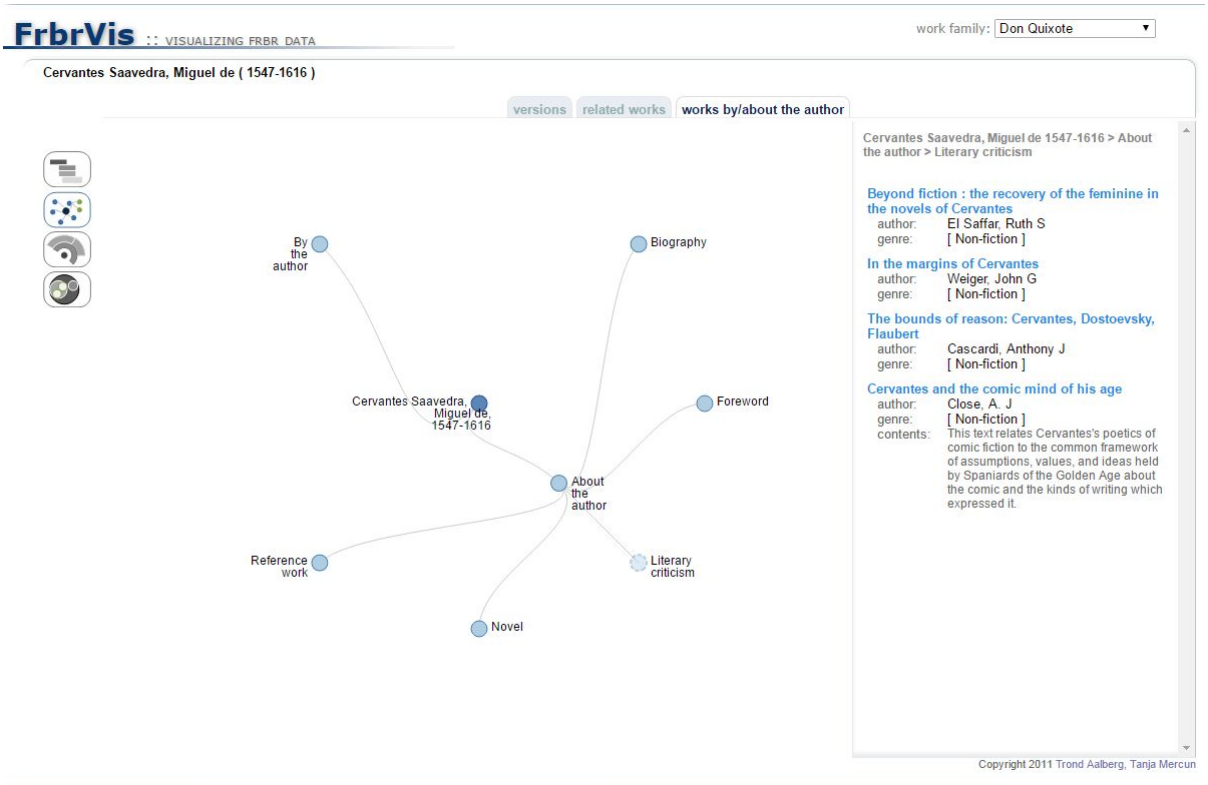


Figure n°2 : Visualisation en graphe hiérarchique de l'arborescence d'une entité auteur exprimée selon le modèle FRBR.[16]

Nous pourrions donc dire que de la même manière que lorsque nous regardons une page de résultats sur le web, nous visualisons à la fois une représentation du réel et l'expression visuelle de la valeur de ce réel attribuée par un langage algorithmique, de la même manière, lorsque nous observons un graphe comme celui de la figure ci-dessus, nous visualisons un langage qui pointe vers des objets réels tout en se représentant lui-même. Nous irions même jusqu'à dire qu'en réalité, ce langage se représente davantage lui-même qu'il ne représente les choses, car on aurait tort de penser que l'algorithme de Google ou que les facettes classificatoires parviennent jamais à retranscrire fidèlement la subjectivité humaine pour le premier, la complexité et l'ambiguïté du monde pour les secondes. Si donc l'on veut pouvoir fournir un accès à l'information qui soit à la fois démocratique, diversifié et original, il est nécessaire de repenser le paradigme épistémologique qui doit déterminer la technologie algorithmique et visuelle à utiliser pour ce faire.

Penser la mise en scène des données du catalogue dans un système de ressemblances

Il paraît trivial aujourd'hui de dire que les bibliothèques, et à plus forte raison leurs données, ont à voir avec le politique : l'environnement du web nous incite de plus en plus à penser ce que pourrait être un accès à l'information dans un cadre démocratique, à une heure où les

[16] MERČUN, Tanja. FrbrVis prototype. http://dijon.idi.ntnu.no/exist/rest/db/frbrvis/index.html (consulté le 27/06/2016).

moteurs de recherche ont tendance à nous enfermer dans une même communauté de pensée et où les classifications dans tous les domaines renverraient en fin de compte qu'à chacun des milieux spécialisés qui les ont vus naître, sans nécessairement faire communiquer ces domaines. Si en premier lieu nous nous intéressons à ce que pourrait être le fait démocratique du point de vue de l'accès à l'information, on s'aperçoit que la sérendipité pourrait en être la pièce maîtresse : en effet, là où nous pourrions définir la démocratie, à la manière de Jean Jaurès, comme la recherche de l'unité dans la diversité, la sérendipité ne se définit pas simplement à ce qu'il nous semble, comme une manière de découvrir par hasard une information que l'on ne cherchait pas au préalable, mais bien aussi comme manière de familiariser chacun avec ce qui est autre que lui-même. De fait, le défi qui se pose aux bibliothèques aujourd'hui ne serait-il pas de rechercher par quel langage sémantique et visuel il serait possible d'éveiller le sujet à son autre ?
De nouveau, la sérendipité, lorsque nous la pensons comme instrument cognitif de découverte de connaissances, nous renvoie vers les ordres de pensée qui ont été ceux du passé, et notamment vers l'encyclopédie d'un certain Tesauro, telle que la décrit Umberto Eco :

> "Le *Cannocchiale aristotelico* d'Emanuele Tesauro (1665), paradoxalement, nous fournit un exemple de modèle encyclopédique. Je dis « paradoxalement » parce que Tesauro, au beau milieu du siècle où s'affirme le modèle de la lunette galiléenne comme instrument paradigmatique pour le développement des sciences naturelles, propose comme instrument de renouvellement de ce que nous appellerions aujourd'hui sciences humaines une lunette qui prend le nom d'Aristote, car l'instrument en question est la métaphore. On reconnaît, dans le *Cannochiale*, le noyau fondamental de la rhétorique aristotélicienne (...), et le modèle de la métaphore y est proposé comme modalité de découverte des relations encore inédites entre les données du savoir – même si, à la différence de Bacon, l'intérêt de Tesauro est d'ordre rhétorique plus que scientifique[17]."

Ainsi, la sérendipité, en tant que possibilité de découverte d'objets nouveaux, ne pourrait-elle pas se transposer dans la métaphore aritsotélicienne, "dont le génie, ajoute Eco, n'est autre que la capacité de « pénétrer les objets fort bien dissimulés sous différentes catégories et de les comparer entre eux », c'est-à-dire la capacité de déceler des analogies et des ressemblances qui seraient passées inaperçues si chaque chose était restée classée dans sa catégorie" ? Or, il est très étonnant d'observer les mêmes présupposés épistémologiques dans les objectifs promus aujourd'hui par les technologies liées à l'exploitation des données massives, notamment dans le domaine de la fouille de données textuelles récemment mis à l'honneur par la numérisation massive de corpus documentaires et leur mise en ligne. C'est ainsi que dans son mémoire intitulé *Big Data et bibliothèques : traitement et analyse informatique des collections numériques*, Johann Gillium décrit les travaux précurseurs de l'informaticien et documentaliste Don R. Swanson dont l'objectif était de "lier entre elles des connaissances disjointes car publiées dans des articles séparées". Ainsi, "selon la formulation canonique de ce type de démarche, si dans un article un lien a été établi entre A et B, et dans un autre un lien entre B et C, il est probable que A influence C également[18]."

Il nous faudrait donc rendre compte d'une certaine dichotomie entre les métadonnées des catalogues et les données textuelles produites en masses par les bibliothèques numériques : alors que les premières nous plongent par la volonté de structuration qui les caractérisent du côté de l'analyse critique et de la fabrication d'une

---

[17] ECO. *Ibidem*, p. 56-57.
[18] GILLIUM, Johann. *Big data et bibliothèques : traitement et analyse informatiques des collections numériques*. Mémoire d'étude Enssib, 2016. p. 23-24.

langue universelle qui ont défini ce que Foucault a appelé l'âge de la représentation, les secondes relèvent davantage de l'exégèse textuelle qui a caractérisé, toujours selon le même Foucault, le XVI[e] siècle et l'âge des ressemblances[19]. Le savoir du XVI[e] siècle s'appuyait en effet sur un langage déjà fait et déposé au préalable, considéré comme brut, à l'opposé du savoir classique dont la volonté était de fabriquer sa propre langue pour y faire refléter le réel comme en un parfait miroir. Or la chaîne des ressemblances est bien là également le présupposé des Big Data, qui, une fois considérés en dehors du mythe de l'objectivité scientifique qui caractérise les discours en faisant la promotion, cherchent à s'appuyer sur un langage de données brut et préexistant dans lequel tout le jeu consistera à chercher le signal faible, le signe qui une fois comparé à un autre signe lui ressemblant, permettra de produire une connaissance nouvelle, un commentaire. Si donc nous pouvons dire que le mode de savoir des données massives est un mode de savoir qui se rapproche davantage de la divination babylonienne, en ce qu'elle repose sur l'idée que le but de la connaissance n'est pas de découvrir les lois fondamentales qui régissent l'univers mais plutôt les liens de corrélations et de significations entre les choses[20], et que de ce point de vue, elle ne peut fonctionner qu'en tant que science d'interprétation et non en tant que science explicative et descriptive, alors, le mode de visualisation fondamental des données massives et notamment des données textuelles est une visualisation dont les buts doivent être les mêmes que ceux de l'astrologie[21]. L'astrologie en effet était par essence une classification sémantique du monde, qui avait son prolongement dans cette visualisation de données que représentait pour les anciens la voûte céleste[22].

> "Les herboristeries astrologiques distinguaient sept plantes planétaires, douze herbes associées aux signes du zodiaque, trente-six plantes attribuées aux décans et aux horoscopes. Les premières, pour être efficaces, devaient être cueillies un certain jour et à une certaine heure, qui étaient précisés pour chacune : dimanche pour le coudrier et l'olivier ; lundi pour la rue, le trèfle,

---

[19] "La Renaissance s'arrêtait devant le fait brut qu'il y avait du langage dans l'épaisseur du monde, un graphisme mêlé aux choses ou courant au-dessous d'elles : des sigles déposés sur les manuscrits ou sur les feuillets des livres. Et toutes ces marques insistantes appelaient un langage second - celui du commentaire et de l'érudition -, pour faire parler et rendre enfin mobile le langage qui sommeillait en elles ; l'être du langage précédait, comme d'un entêtement muet, ce qu'on pouvait lire en lui et les paroles dont on le faisait raisonner." FOUCAULT, *Ibidem*. p. 93.

[20] "Our own natural sciences are based on a premise so simple that it is usually taken for granted : Things behave according to universally valid laws. It is our task to discover those laws, and the means to do so is observation, supported by the controlled experiment. In a similar fashion, Babylonian divination is based on a very simple proposition : Things in the universe relate to one another. Any event, however small, has one or more correlates somewhere else in the world. This was revealed to us in days of yore by the gods, and our task is to refine and expand that body by observation. There is no evidence that the Mesopotamian scholars ever attempted to verify the results of their speculations by experiment. Nevertheless, the Neo-Assyrian astrologers undoubtedly believed in their craft and found it confirmed by events." KOCH, Ulla Susanne. *Mesopotamian astrology : an introduction to Babylonian and Assyrian celestial divination*. Copenhague : Museum Tusculanum Press, 1995. p. 18-19.

[21] Ce parallèle entre Big Data et astrologie a d'ailleurs été établi à l'occasion d'une étude statistique du Columbia University Medical Center : GIBBS, Mark "Big Data and astrology? Your health correlates with the month you were born!". http://www.networkworld.com/article/2984688/big-data-business-intelligence/big-data-and-astrology-your-health-correlates-to-the-month-you-were-born.html (consulté le 27/06/2016).

[22] "J'ai montré ailleurs que l'on pouvait analyser le zodiaque astrologique comme l'une des composantes de ce que Lévi-Strauss appelle dans *La Pensée Sauvage* un système de transformations. Avec les références aux directions cardinales qui le constituent (équinoxe et solstices), les combinaisons d'éléments qui caractérisent les signes (eau, air, terre, feu), les dénominations de ces derniers et les propriétés qui s'y associent, enfin leur stricte alternance binaire en masculins et féminins ou diurnes et nocturnes, le zodiaque fait partie d'un système complexe de classification, permettant d'encoder tous les phénomènes naturels et de transformer certains d'entre eux en symboles." SIMON, Gérard. *Sciences et savoirs aux XVI[e] et XVII[e] siècles*. Lille : Presses du Septentrion, 1996. p. 72.

> la pivoine, la chicorée ; mardi, pour la verveine ; mercredi pour la pervenche, la pivoine, le cytise et la quintefeuille si on les destine à des usages médicinaux ; le vendredi pour la chicorée, la mandragore et la verveine servant aux incantations ; samedi, pour la cruciata et le plantain. On trouve même chez Théophraste un système de correspondances entre les plantes et les oiseaux, où la pivoine est associée au pic, la centauride au triorchis et au faucon, l'ellébore noir à l'aigle (...).[23]"

Par son principe visuel, symbolique et sémantique, ce type de pensée analogique pourrait peut-être répondre à la question qui se pose immédiatement chaque fois que l'on s'essaie à la tâche de visualiser des données massives, à savoir le moyen par lequel on pourrait rendre compréhensible, navigable et propice à la sérendipité la quantité illimitée de données offerte par le numérique. D'une certaine manière, la pensée "totémique" permettrait peut-être d'accorder la masse des données à notre disposition avec la rareté de l'attention qui prévaut dans l'environnement du web. Dans le domaine des bibliothèques, un exemple de visualisation de corpus par ressemblances peut-être vu dans le prototype en ligne "The Bohemian Bookshelf"[24] (voir figure n°3 ci-dessous) : les ouvrages y sont à la fois visuellement représentés et classés en fonction de leurs ressemblances physiques (nombre de pages et couleur de la couverture) mais également en fonction de leurs proximités de contenus (sujets liés, ordre alphabétique des auteurs et temporalité des ouvrages).

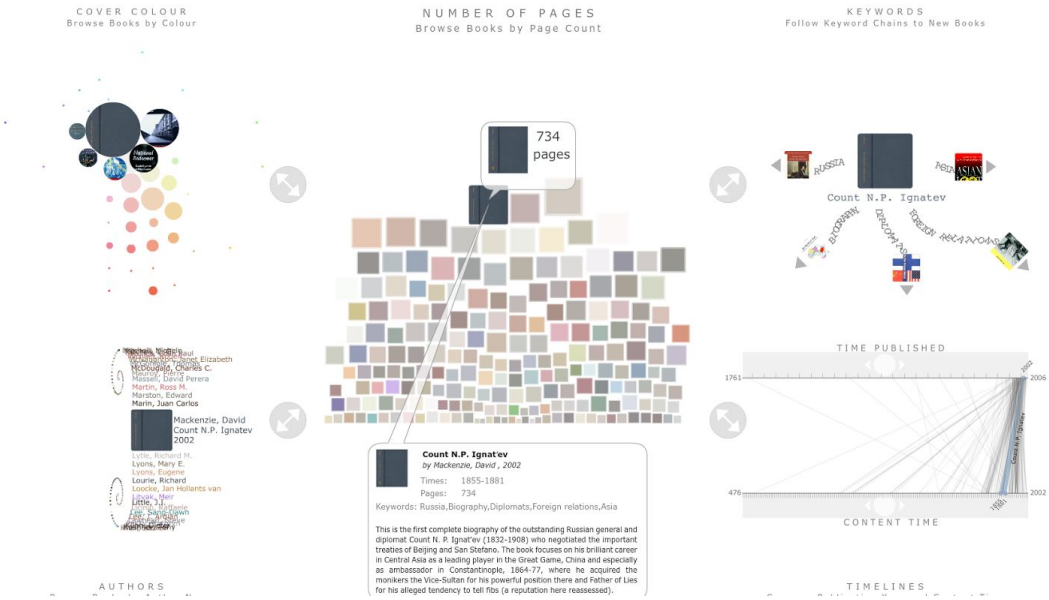


Figure n°3 : "The Bohemian Bookshelf", par Alice Thudt.

Peut-être s'agirait-il désormais, pour les catalogues des bibliothèques, de s'appuyer sur le web sémantique et sa capacité à lier et à nommer pour promouvoir une navigation par ressemblances métaphoriques au sein des corpus documentaires qu'ils décrivent : donner une épaisseur aux liens entre des données d'auteur, d'oeuvre, de sujet, etc., en faisant de ces derniers une donnée à part entière constituée automatiquement à partir de corrélations et de recoupements. Ainsi, d'une certaine manière, les données massives et la fouille de données textuelles feront-elles peut-être renouer les bibliothèques avec leur passé lointain, et notamment avec les modèles antiques de bibliothèque astronomique qu'étaient la bibliothèque royale d'Assurbanipal et la bibliothèque d'Alexandrie.

---

[23] LEVI-STRAUSS, Claude. *La Pensée Sauvage*, Paris : Plon, 1962. p. 58.
[24] THUDT, Alice. "The Bohemian Bookshelf".
http://www.alicethudt.de/BohemianBookshelf/Program/BB.swf (consulté le 03/06/2016).

Conclusion

Le web est un théâtre de signes, une scène où se succèdent et s'organisent dans un ordre déterminé au préalable des représentations du monde et de la valeur que leur attribuent pour nous des acteurs institutionnels dont les visions économiques, certes opposées, ne sont pour autant pas imperméables les unes des autres : il y a bien de discrètes facettes classificatoires sur les pages de résultats des moteurs de recherche, et si le moteur de recherche de wikipédia n'est plus guère utilisé par les internautes, il n'en n'a pas pour autant été effacé des pages de l'encyclopédie en ligne. Il s'agit en effet d'affirmer ici l'appartenance de ces deux visions, "utilitariste" et "physiocrate", à un même segment de pensée, ainsi que leur disposition à un extrême ou l'autre de ce segment, sans séparation ni discontinuité : il est davantage question pour nous de définir des centres de gravité plutôt que de circonscrire avec certitude le comportement des différents acteurs du web dont d'une certaine manière le catalogue des bibliothèques, par son principe même, a toujours été partie intégrante.

Le web est donc un théâtre de signes, mais de signes qui, en définitive, ne parviennent qu'à représenter leur fonction de représentation avant de représenter le réel. Ainsi, le prédicat, cheville ouvrière du web sémantique, représente en lui-même les deux principales fonctions de classification et d'attribution qui sont chacune au coeur des logiques du versant économique de la représentation : on pourrait donc dire que d'une certaine manière, le prédicat est emblématique de la logique représentationnelle du web. Or, le fonctionnement de ce prédicat est très comparable à celui qu'attribuaient les grammairiens de l'Ancien Régime au verbe dans le langage dit "classique" : "toile invisible, entièrement recouverte par le dessin des mots, mais qui donne au langage le lieu où faire valoir sa peinture ; ce que le verbe désigne, c'est finalement le caractère représentatif du langage, le fait qu'il ait son lieu dans la pensée, et que le seul mot qui puisse franchir la limite des signes et les fonder en vérité, n'atteigne jamais que la représentation elle-même"[25]. Si donc le verbe, et en particulier le verbe être était pour le "classicisme" la toile où se peignait le monde, de même, le prédicat construit-il dans le lien hypertexte la "grande toile" sur laquelle s'appuient les différents acteurs du web, catalogues compris, pour mettre en scène leur propre représentation du monde.

Les données massives et l'ordre du savoir exégétique et non critique qui est le leur, il est vrai, bouleversent quelque peu cette logique du web. Encore faudrait-il les considérer pour ce qu'elles sont, à savoir non une science du fait objectif mais un art du signe et de l'interprétation. Peut-être serait-il nécessaire, pour en arriver là, de s'efforcer de penser le média du web et du catalogue sur le web du point de vue d'une certaine transparence démocratique, c'est-à-dire comme un regard porté non sur les données et leur signifiant mais bien sur celui qui produit la donnée et son signifiant. La sérendipité, en tant qu'ouverture sur l'altérité, ne peux véritablement être trouvée que lorsque le langage des données aura fini de ne signifier que lui-même et commencé de renvoyer à son propre mode d'être, ses codes et son cadre de production. "Dans notre art théâtral, écrit Roland Barthes, l'acteur feint d'agir, mais ses actes ne sont jamais que des gestes : sur la scène, rien que du

[25] FOUCAULT. *Ibidem*, p. 110.

théâtre, et cependant du théâtre honteux. Le *Bunraku*[26], lui, (c'est sa définition), sépare l'acte du geste : il montre le geste, il laisse voir l'acte, il expose à la fois l'art et le travail, réserve à chacun d'eux son écriture."[27]

[26] Théâtre de poupées japonais.
[27] BARTHES, Roland, 2002. *Oeuvres complètes. Tome 3, 1968-1971*. Seuil. p. 394.