Diplôme de Conservateur des Bibliothèques

Mémoire de fin d'étude / décembre 2014

Faire parler les données des bibliothèques : du Big Data à la visualisation de données

Raphaëlle Lapôtre

Sous la direction de Julien Velcin
Maître de Conférence en informatique – Université Lumière Lyon 2

Remerciements

Mes remerciements vont en premier lieu à mon directeur de mémoire, Julien Velcin pour sa patience, son dévouement et sa compréhension : ce travail est largement redevable tant aux précieux conseils qu'il m'a prodigués qu'à la liberté d'actions qu'il m'a laissée tout au long de sa rédaction. Ils s'adressent ensuite à Bertrand Calenge, qui a bien voulu faire confiance à ce projet pourtant complexe et qui l'a accompagné de sa bienveillance tout au long de son élaboration. Je remercie également Véronique Poirier, Jean-Pierre Berthon, Valérie Bouissou et Denis Cordazzo pour l'enthousiasme qu'ils avaient manifesté lors de mes premières et modestes expériences avec les données de la Bibliothèque Publique d'Information. Enfin, un grand merci à Florent Derex, Dominique et Didier Lapôtre, Morgane Spinec, Louise Daguet et Marc Bruchet qui ont supporté quotidiennement mes états d'âmes et mes doutes depuis le premier jour de cette entreprise.

Résumé : Cette étude se penche sur les enjeux de la réutilisation des données des bibliothèques à l'ère du Big Data. En ce qui concerne la production de connaissances sur le monde des bibliothèques et de l'information, les technologies d'analyse du Big Data, contrairement à ce que prétendent les discours qui peuvent parfois les accompagner, ne réduisent pas les biais et présupposés inhérents aux statistiques traditionnelles. Cependant, la visualisation de données, telle que revue et critiquée par les Humanités Numériques, pourrait permettre de prendre en compte d'une manière beaucoup plus centrale la nature fondamentalement politique des bibliothèques. Regardant le pilotage des établissements documentaires, certains auteurs appellent à fonder les décisions non sur les données et chiffres mais sur l'analyse de données. De fait, l'ouverture de la profession de bibliothécaire sur la science des données pourrait être un bon moyen de faire évoluer les méthodes d'évaluation et de pilotage. La visualisation est un moyen ludique d'apprendre l'analyse de donnée et permet de communiquer efficacement sur l'activité de l'établissement. En dernier lieu, les discours actuels accompagnant l'ère du numérique font l'apologie d'un accès individualisé et fragmenté à l'information qui permettrait de se passer des biais inhérents à toute classification universelle. Néanmoins, ces biais sont transposé dans les algorithmes de recherche de l'information. Dès lors, il devient nécessaire de penser un système de navigation qui exprime ce biais et le soumette davantage à une discussion : transformer un catalogue de bibliothèque en data game pourrait être une solution pour exprimer de manière ludique la métaphore sous-jacente à toute organisation des connaissances.

Descripteurs : Big Data, visualisation, interface de navigation, classification, métaphore, évaluation, communication, Patron-Driven Acquisition.

Abstract : This work is about the issues raised by the re-use of library data at the age of Big Data. Regarding the production of knowledge about libraries and their users, the new analysis technologies are not reducing inherent bias of traditional statistics. Nevertheless, data visualization as considered by the Digital Humanities is a very interesting tool, because it make the human subjectivity implied by such technologie a central element through which we can consider the library more as a political object.

As for library management, authors are calling for analysis-driven rather than data-driven decisions. Thus, training librarians in data analysis could be a good solution, in the context of open data and open research data. Data visualization is a funny way to learn data analysis and is a very effective way of communicating about the library activities. Lastly, if it can be read that access to information at the digital age is now more individual and can allow to circumvent the bias of traditional classification, we claim that those bias are transposed in the algorithms that allow this access today. Thus, it is important to consider a way of navigating into the information that make obvious and submit for discussion those bias. In this respect, a library catalog conceived as a data game is a metaphoric and funny way to explore library collections while not taking too seriously such an knowledge organisation.

Keywords : Big Data, library data, data visualization, Patron Driven Acquisition, Evaluation, Communication, browsing interface, metaphore, classification.

Droits d'auteurs

Sommaire

Sigles et abréviations

ADS : Astrophysics Data System
API : Application Programming Interface
BIUSJ : Bibliothèque Interuniversitaire Scientifique de Jussieu
BnF : Bibliothèque Nationale de France
Bpi : Bibliothèque Publique d'Information
BUPMC : Bibliothèque Universitaire Pierre et Marie Curie
CSV : Comma Separated Values
DPLA : Digital Public Library of America
DST4L : Data Scientist Training For Librarians
EMEA : Europe Middle East Africa
EVR : Extension Visuelle de Requête
FRBR : Functional Requirement for Bibliographic Records
JPEG : Joint Photographic Expert Group
JSON : Javascript Object Notation
K-NN : k-Nearest Neighbors
NASA : National Aeronautics and Space Agency
OCLC : Online Computer Library Center
OPAC : Online Public Access Catalog
PDF : Portable Document Format
PEB : Prêt Entre Bibliothèques
SICD : Service Interétablissement de Coopération Documentaire
VIAF : Virtual International Authority File
XML : Extensible Markup Language

INTRODUCTION

De mars à juin 2013, j'ai eu l'opportunité d'effectuer un stage d'observation à la Bibliothèque Publique d'Information, à l'occasion duquel Véronique Poirier, déléguée à la politique documentaire de l'établissement, m'avait demandé de réfléchir à l'approfondissement des critères de désherbage afin de pouvoir éliminer de manière plus importante certains ouvrages imprimés, notamment dans les domaines cumulatifs que sont les lettres et sciences humaines. En pleine réorganisation, l'établissement envisageait en effet le déménagement de ses collections et se posait avec acuité la question de la saturation de certaines de ses étagères, la Bpi ne disposant pas de magasins permettant de stocker le surplus éventuel de ses collections. Dans un premier temps, j'ai donc exploré le catalogue en ligne en espérant assez naïvement repérer quelques zones sensibles, des endroits où la collection était peut-être trop ancienne (si tant est que ce critère puisse être valable en lettre et sciences humaines), peut-être trop spécialisée par rapport au grand public de la Bpi (mais de ce point de vue, une thèse d'université peut parfois se révéler plus accessible qu'un manuel de premier cycle universitaire), peut-être trop redondante dans les sujets couverts (mais certains sujets sont naturellement plus couverts que d'autres)... Finalement découragée par la masse des collections a explorer, je me suis décidée à recourir à l'entrepôt de données de la Bpi afin d'extraire des listes d'ouvrages par domaine et d'essayer de les synthétiser par des estimateurs statistiques : il s'agissait de déterminer, pour chaque tranche de cote, les dates d'édition les plus anciennes, les plus récentes, l'âge moyen de l'ensemble, les auteurs et les sujets les plus représentés, en essayant de visualiser cela au moyen de diagrammes en barres et de simples « camemberts ». Naturellement, je ne pouvais pas déduire grand chose de ces analyses exploratoires sans une connaissance approfondies des réalités documentaires vers lesquelles elles pointaient, et j'ai donc pris la décision d'interroger des responsables de collections sur mon petit travail statistique. Ce fut sans doute la partie la plus intéressante de mon stage : au fil des conversations, des réactions de tous ordres sur les spécificités que dessinaient moyennes, médianes, maximums et minimums, se laissaient entrevoir l'histoire des collections de la BPI, la personnalité des personnes qui avaient contribué à les façonner années après années, les événements qui avaient pu marquer un changement dans la manière de collecter, la proximité marquée du Centre Georges Pompidou, les visions du monde différentes que pouvaient révéler les conflits dans les manières de classer les ouvrages... La richesse culturelle et institutionnelle de la Bpi émergeait ainsi, me semble-t-il, de ces aller-retours entre les données et les conversations avec les professionnels qui acceptaient de réagir à ce que je leur montrais.

Si je pourrais difficilement qualifier ma contribution au désherbage des collections de la BPI de décisive, ce stage a cependant aiguisé ma curiosité pour les données, et notamment les données des bibliothèques : comment en effet définir ces dernières ? Pourquoi serait-il particulièrement intéressant d'en parler en 2014 ? Comment faire parler les données des bibliothèques et, surtout, dans quel but ? Ces questions, dont les réponses ne peuvent être définitives, me semblent cependant constituer un préalable essentiel à une étude qui voudrait essayer de percevoir toutes les possibilités et les limites offertes par la réutilisation des données des bibliothèques. C'est ainsi que, pour tenter d'y répondre, il me faudra passer du « je » de l'expérience personnelle au « nous » du mémoire d'étude.

Qu'est-ce que les données des bibliothèques ?

Pour parler des données des bibliothèques nous pourrions peut-être dans un premier temps nous pencher sur la question d'une définition plus générale de la « donnée ». Si l'on reprend en effet les mots de Lynda Kellam et Katharin Peter, les données seraient « toute information structurée d'une manière reconnaissable »[1], incluant donc à la fois des informations qualitatives et quantitatives dont le point commun est d'avoir été collectées, traitées et organisées de façon à les rendre compréhensibles. Lorsque l'on parle de données cependant, notamment dans le domaine de l'administration et des bibliothèques, il arrive souvent que l'on ne désigne sous ce terme que des chiffres, qu'ils désignent des quantités de quelque chose ou des agrégats de ces quantités produits par synthèse statistique, comme le sont les totaux, pourcentages, moyennes et autres médianes... La plupart du temps, ces chiffres se laissent entrevoir sous forme de tableaux et de graphiques, le tableau pouvant être considéré comme une forme de visualisation des données au même titre qu'un graphique. D'une certaine manière, il est donc difficile de parler des données sans se représenter les logiciels les plus communs qui permettent de les traiter, à savoir Excel, SPSS[2] ou les bases de données relationnelles telles que MySQL. Dans ce contexte, c'est la notion de « jeu de données » qui se profile, à savoir, selon Rémi Gaillard, « l'agrégation, sous une forme lisible, de données brutes ou dérivées présentant une certaine « unité », rassemblées pour former un ensemble cohérent »[3]. Dès lors, les données brutes désignent des « microdonnées », à savoir un unique enregistrement de quelque chose, et les données dérivées, des données produites à partir d'un premier jeu par nettoyage et synthèse statistique.

À ce stade de notre réflexion, il nous semble important de souligner le caractère ambiguë de la « donnée » : là où en effet Rémi Gaillard affirmait qu'une donnée pouvait rarement être isolée de son contexte de production, et de ce fait ne pouvait être qualifiée de « brute » qu'avec une certaine précaution[4], nous aimerions ajouter avec Johanna Drucker qu'il serait préférable de parler non de « donnée » mais de « captée »[5] : la donnée n'est pas un objet produit d'une manière extérieure et indépendante de l'homme par le réel, elle est d'avantage une perception enregistrée du réel, construite selon certaines techniques et certaines contraintes, sélectionnée pour répondre à des objectifs définis en amont de sa conception.

Comment définir, dans ce contexte, les données des bibliothèques ? Celles qui viennent immédiatement à l'esprit sont les métadonnées, à savoir les données décrivant les documents de la bibliothèque. Or, le caractère éminemment ontologique de ces métadonnées doit être souligné : c'est en effet cet aspect qui fait la richesse des données des bibliothèques, et par extension, des données des institutions culturelles (musées, archives, arboretum, etc.). Là où en effet les autres données peuvent être généralement considérées comme de simple reflets du réels, les métadonnées pointent directement vers lui en posant la question de la nature de

[1] KELLAM, Lynda M et PETER, Katharin, 2011. *Numeric data services and sources for the general reference librarian*. Oxford : Chandos Publishing. p. 7-8.

[2] « SPSS (Statistical Package for the Social Sciences) est un logiciel utilisé pour l'analyse statistique. C'est aussi le nom de la société qui le revend (SPSS Inc). » SPSS, 2014. *Wikipédia* [en ligne]. [Consulté le 12 décembre 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=SPSS&oldid=109086133. Page Version ID: 109086133

[3] GAILLARD, Rémi, 2013. *De l'Open data à l'Open research data quelle(s) politique(s) pour les données de recherche ?* Bibliothèque Numérique de l'Enssib. Consulté le 18 août 2014. Disponible à l'adresse Web : http://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf p. 19.

[4] *Ibid*. p. 18.

[5] DRUCKER, Johanna, 2011. Humanities Approaches to Graphical Display. [en ligne]. 2011. Vol. 5, n° 1. [Consulté le 1 novembre 2014]. Disponible à l'adresse : http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

à quoi elles font référence[6]. Mais en dehors de ces métadonnées, existent pour les bibliothèques les données relatives à leurs activités : acquisitions, désherbages, jauge de fréquentation, circulations des documents, inventaire des collections, données de logs captées au moment où les utilisateurs se connectent au site internet de la bibliothèque ou bien à ses bases de données. Si nous devions donner un exemple précis de ces données, peut-être pourrions-nous décrire un tableau contenant des informations relatives aux abonnements de la bibliothèque à des périodiques. Ainsi pourrait-on y lire des informations sur le mode d'acquisition (abonnement, don, dépôt légal), le nombre de numéros réguliers à recevoir, la date d'arrivée prévue du premier numéro, la date d'annulation, la date de la facture, la date de parution du dernier numéro généré, le numéro d'identification unique du titre, le numéro logique de la notice bibliographique, etc. Un tel tableau contenant un ensemble de variables relatives à plusieurs objets de nature similaire est ce qu'on appelle un jeu de donnée[7].

Pourquoi parler des données des bibliothèques en 2014 ?

La « révolution » du Big Data, ou mégadonnées est toujours au cœur de l'actualité en 2014, en témoigne la sortie, le 20 février dernier, de la traduction française du best-seller de Kenneth Cukier et Victor Mayer-Schoenberger sur le sujet[8]. Force est de constater que la notion de Big Data recouvre malgré cela une réalité bien difficile à définir. Selon le Wikipédia anglais, celle-ci désigne un ensemble de processus de traitement de jeux de données dont le volume est tel qu'il n'est plus possible d'employer les méthodes traditionnelles pour les traiter[9]. Cependant, la taille des données du Big Data est toute relative, comme l'expliquent Cathy O'Neil et Rachel Schutt dans leur ouvrage de référence, *Doing Data Science*[10] : en réalité, le big de Big Data peut désigner un petabyte au même titre qu'un terabyte ou un gigabyte... seul importerait véritablement le fait que la masse de donnée dépasse la capacité de stockage et de vitesse de traitement des machines actuelles, cette capacité évoluant quant à elle avec son époque. Plus concrètement, si nous devions rester dans la perspective choisie pour cette étude, à savoir la réutilisation des données notamment dans le but de connaître et de piloter un ou plusieurs établissements, il conviendrait peut-être de retenir la définition fournie par Steve Lohr dans un article du New York Time datant de 2013[11], qui présente le Big Data comme une nouvelle manière de prendre des décisions en se fondant sur l'analyse de grandes masses de données telles qu'elle est rendue possible par les technologies d'aujourd'hui. Malgré cela, il est nécessaire de rappeler que le mouvement des mégadonnées ne se réduit pas à la simple prise de décision informée par les données : il se caractérise également par une volonté de pérenniser les données et d'en extraire des connaissances si possible nouvelles, sans toutefois que cette connaissance soit subordonnée à l'aspect politique et décisionnel sous l'angle duquel la question est examinée dans ce mémoire.

---

[6] The Life and Death of Data, [sans date]. [en ligne]. [Consulté le 2 novembre 2014]. Disponible à l'adresse : http://lifeanddeathofdata.org/

[7] Data set, 2014. *Wikipedia, the free encyclopedia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://en.wikipedia.org/w/index.php?title=Data_set&oldid=625099781. Page Version ID: 625099781

[8] CUKIER, Kenneth, MAYER-SCHOENBERGER, Viktor et DHIFALLAH, Hayet, 2014. *Big Data*. Paris : ROBERT LAFFONT.

[9] Big data, 2014. *Wikipedia, the free encyclopedia* [en ligne]. [Consulté le 1 novembre 2014]. Disponible à l'adresse : http://en.wikipedia.org/w/index.php?title=Big_data&oldid=631791921. Page Version ID: 631791921

[10] O'NEIL, Cathy, SCHUTT, Rachel. *Doing Data Science*, [sans date]. [en ligne]. [Consulté le 1 novembre 2014]. Disponible à l'adresse : http://shop.oreilly.com/product/0636920028529.do. Non paginé dans sa version électronique.

[11] Sizing Up Big Data, Broadening Beyond the Internet, [sans date]. *Bits Blog* [en ligne]. [Consulté le 1 novembre 2014]. Disponible à l'adresse : http://bits.blogs.nytimes.com/2013/06/19/sizing-up-big-data-broadening-beyond-the-internet/. « Big Data is a vague term, used looely, if often, these days. But put simply, the catchall phrase means three things. First it's a bundle of technologies. Second it's a potential revolution in measurement. And third, it is a point of view, or philosophy, about how decisions will be – and perhaps should be – made in the future ».

D'une certaine manière, nous pourrions dire que les bibliothèques contribuent à y participer par la masse traditionnellement importante de leur propres données, mais aussi par la volonté d'ouverture qui les accompagne : le mouvement des mégadonnées intervient en effet dans un contexte d'échange et de libre partage des données, sans obstacle juridique, technique ou financier, et c'est là précisément la définition de l'open data. Or, les bibliothèques sont directement concernées par l'ouverture de leurs données, notamment de leurs données bibliographiques : les initiatives comme celle de la BnF[12], visant à faire des bibliothèques des acteurs du web de données en exposant leurs données sur le web et en les reliant entre elles, participent à ce mouvement d'ouverture en favorisant la réutilisation à grande échelle de ces données.

A cette ouverture de certaines des données des bibliothèques s'ajoute la participation des bibliothèques universitaires au mouvement d'ouverture des données de la recherche : afin de rendre possible une transparence et une communication plus grande des méthodes scientifiques et des données produites dans le contexte de la recherche, ces dernières sont de plus en plus incitées à mettre à disposition de leurs usagers des dispositifs de stockage et de réutilisation de ces données. Nous aimerions souligner ici que ce mouvement ne concerne pas seulement les bibliothèques de recherche, mais peut également toucher les bibliothèques publiques. Par ailleurs, il ne concerne pas seulement les sciences dures mais aussi les lettres et sciences sociales : le mouvement des humanités numériques, dont les techniques ont inspiré une partie de cette étude, s'appuie en effet massivement sur les nouvelles possibilités de stockage et de traitement des données offertes par les technologies en 2014.

Comment faire parler les données ?

En premier lieu, qu'entend-t-on par « faire parler » les données ? Une première réponse à cette question peut se trouver dans la définition que donnent Cathy O'Neil et Rachel Schutt à l'inférence statistique, à savoir « la discipline qui se préoccupe du développement de procédures, de méthodes et de théorèmes qui nous permettent d'extraire du sens et de l'information de données qui ont été générées par un processus stochastique (aléatoire) »[13]. Faire parler les données, ce serait donc en premier lieu en « extraire du sens et de l'information », tout en sachant que ce sens est d'avantage construit qu'extrait. Plus précisément, l'inférence statistique désigne des déductions produites sur une population à partir d'un échantillon de cette population que l'on observe. Mais en 2014, les acteurs du Big Data se contentent-ils de statistiques inférentielles pour faire parler leurs données ? De fait, ce procédé s'inscrit dans la discipline plus générale de ce qu'on appelle aujourd'hui la « science[14] des données », suivant les propos d'O'Neil et Schutt : « plus précisément, un scientifique des données est une personne qui sait comment extraire du sens des données et les interpréter, ce qui nécessite à la fois des outils et des méthodes provenant des statistiques et de l'apprentissage automatique, et aussi d'être humain »[15].

[12] FRANCE, Bibliothèque nationale de, [sans date]. BnF - Les enjeux du web de données en bibliothèque. [en ligne]. [Consulté le 2 novembre 2014]. Disponible à l'adresse : http://www.bnf.fr/fr/professionnels/innov_num_web_donnees/a.web_donnees_enjeux_bibliotheques.html

[13] O'NEIL, SCHUTT, 2013. « More precisely, statistical inference is he discipline that concerns itself with the development of procedures, methods, and theorems that allow us to extract meaning and information from data that has been generaed by stochastic (random) processes ». Non paginé dans sa version électronique.

[14] L'appellation de « science » pour cette discipline fait aujourd'hui débat : O'Neil et Schutt y voient plutôt un art.

[15] *Ibid*. « More generally, a data scientist is someone who knows how to extract meaning from and interpret data, which requires both tools and methods from statistics and machine learning as well as being human ». Non paginé dans sa version électronique.

Mais s'il semble évident que la science des données comprendrait l'ensemble des méthodes visant à faire parler les données, la définition de cette « science » n'en reste pas moins délicate à cerner. Peut-être pourrait-on commencer par l'exemple qui illustre le mieux ce procédé dans le monde de l'information et des bibliothèques : celui d'un système de recommandation. La construction de ce type de système nécessite en effet de savoir utiliser un large éventail de méthodes relevant de la science des données : il faut commencer par créer un réseau, ou graphe, entre des données décrivant des utilisateurs (par exemple, des lecteurs), et des données de produits (par exemple, des livres). Il faut ensuite apprendre à un ordinateur à regrouper des lecteurs et des livres en fonction de leur préférence, en s'inspirant de préférences déjà exprimées par le passé. Ce classement s'appuiera lui-même sur un algorithme, c'est-à-dire un mode d'emploi permettant d'accomplir une tâche particulière : en l'occurrence, cela pourrait l'être l'algorithme des plus proches voisins (ou k-NN, pour k-Nearest Neighborgs), dont le but est de classer un ensemble d'objets à partir d'un classement qui a déjà été effectué sur des objets similaires. En dernier lieu, la particularité d'un système de recommandation est qu'il crée une boucle de rétroaction, au sens où son utilisation sur le web pourra influencer des utilisateurs qui, en retour, influenceront le système de recommandation grâce aux données générées par leur comportement.

Mise en réseau, apprentissage automatique, algorithmes sont des moyens de faire parler les données, mais il en existe un autre sur lequel les humanités numérique s'appuient particulièrement, à savoir la visualisation des données. Si l'on devait définir assez généralement cette dernière, nous pourrions retenir les termes de Wikipédia à savoir une « représentation graphique de données statistiques », fournissant un « résumé visuel des données statistiques chiffrées » et permettant de saisir en un seul coup d'oeil « la tendance générale »[16]. Néanmoins, la visualisation dont nous parlerons dans la suite de cette étude prendra pour une grande partie sa référence dans l'usage qu'en font les Humanités Numériques, davantage que dans la perception qu'en ont les scientifiques des données. Prenons ainsi la définition donnée par l'auteur de l'essai intitulé « The life and death of metadata »[17], mis en ligne dans le cadre du Metalab[18], un laboratoire d'Harvard fondé par Jeffrey Schnapp et dédié à la « culture en réseau » :

> « (…) Je propose de penser la visualisation de données comme des « projections », pour souligner la qualité spéculative de telles images ainsi que leur lien avec la pensée métaphorique. (…) En effet, les visualisations ne sont pas autre chose que des métaphores visuelles, transposant divers types de données quantitatives sous forme graphique et spatiale. En tant que métaphores, les visualisations relient des domaines source (des jeux de données) à des domaines cible (des structures graphiques. Par exemple, dans les visualisations temporelles décrite dans cet essai, le temps (à partir des données d'accès) est relié à l'espace de l'écran (en coordonnées) »[19].

La visualisation serait donc ce qui permet de projeter spatialement les données afin de mettre en évidence de manière directe les tendances et particularités que ces données prises dans leur ensemble sont susceptibles de manifester. Il convient de rebondir ici sur

[16] Représentation graphique de données statistiques, 2014. *Wikipédia* [en ligne]. [Consulté le 12 décembre 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Repr%C3%A9sentation_graphique_de_donn%C3%A9es_statistiques&oldid=108854835. Page Version ID: 108854835

[17] The Life and Death of Data, [sans date].

[18] About | metaLAB (at) Harvard, [sans date]. [en ligne]. [Consulté le 7 août 2014]. Disponible à l'adresse : http://metalab.harvard.edu/about/

[19] The Life and Death of Data. [sans date]. « (…) I propose thinking of data visualizations as « projections », to emphasize the speculative quality of such images as well as their relationship to metaphorical thinking. (…) Indeed, visualizations are no more than visual metaphors, translating various kinds of quantitative data into spatial and graphical form. As metaphors, visualizations map source domains (data sets) to target domains (graphical structures). For instance, in the timeline visualizations portrayed in this essay, time (from accession data) is mapped onto the space of the screen (in coordinates) ».

le terme « manifester » : la raison pour laquelle nous avons choisi comme fil directeur de cette étude la visualisation est cette qualité intrinsèque qu'elle peut avoir, lorsque elle est utilisée dans le cadre des humanités numériques, à reconnaître explicitement son caractère construit et les présupposés sur lesquels elle s'appuie. Par opposition, les autres techniques dont nous avons parlé plus haut auraient tendance à considérer que les techniques d'interprétation des données prolongent l'humain. Cependant, elles ne font pas de ce dernier une caractéristique centrale du processus d'interprétation des données, ce qui peut être préjudiciable lorsqu'il s'agit de faire parler les données provenant de cet objet humain, social et politique qu'est la bibliothèque, et plus généralement, l'information. On ne saurait donc écrire que les données manifestent une connaissance : nous *faisons en sorte qu*'elles la manifeste dans toute les phases de leur élaboration, depuis leur collection jusqu'à leur organisation en un ensemble structuré et lisible à la fois par un ordinateur et par une personne.

Ayant conscience du caractère obscur de cette dernière proposition, cette étude tentera d'approfondir cet aspect particulier mais central, et pour lequel il nous semble qu'il vaille véritablement la peine de s'intéresser aux données des bibliothèques, à savoir la nature construite et artificielle du sens et du discours que l'on peut faire émerger de ces données sur les institutions qui les ont produites. Nous proposons donc d'éclaircir cela en nous penchant sur trois domaines dans lesquels les bibliothèques ont traditionnellement utilisé les données.

Tout d'abord, un premier domaine qui est, concédons-le, assez vague et large, puisqu'il va de l'étude de l'évolution des pratiques culturelles – et notamment de la lecture et de la fréquentation des bibliothèques –, à l'histoire des bibliothèques elles-mêmes : s'il est vrai que les nouvelles techniques apportées par la science des données sont susceptibles de bouleverser les méthodes traditionnelles qui nous permettaient de produire des connaissances sur les bibliothèques et leur public, il reste que la méthode la plus honnête et la plus probante à nos yeux est celle de la visualisation, et notamment la visualisation des métadonnées : par leur caractère d'artefact, les métadonnées révèlent les conditions matérielles, les systèmes logiques et classificatoires, les valeurs institutionnelles et culturelles qui les ont vu naître.

Nous nous penchons ensuite sur les enjeux soulevés par l'application de la science des données au pilotage des établissements documentaires. À une époque où la tendance consiste à fonder les décisions sur des preuves chiffrées, peut-être serait-il bon de redonner du sens à l'utilisation des données. Ainsi, par son caractère métaphorique et ludique, la visualisation peut-elle être une bonne approche pour apprendre à manipuler les données tout en permettant une communication efficace sur l'établissement et son activité. Cet apprentissage peut par ailleurs s'appuyer sur ce vaste mouvement qui tend à ouvrir les compétences des bibliothécaires sur la culture des données, notamment dans le cadre de la recherche.

Enfin, un troisième domaine d'utilisation des données des bibliothèques est celui des catalogues en ligne : les OPAC, au même titre que les systèmes de recommandation, sont l'exemple type d'un « produit de données », à savoir un dispositif permettant à un acteur d'interagir avec son public, cette l'interaction pouvant être utilisée pour modifier ce dispositif. De manière plus concrète, les catalogues en ligne permettent au public d'une bibliothèque d'explorer virtuellement sa collection, chaque donnée d'une notice étant à sa manière une métaphore pointant vers la réalité physique du livre qu'elle désigne et qui se trouve localisé dans la bibliothèque. La visualisation nous paraît être une méthode

permettant de proposer une exploration collective de la connaissance tout en soulignant par son caractère métaphorique la faillibilité de cette proposition. Elle demeure sans doute un moyen intéressant d'explorer virtuellement les collections et d'animer une communauté de lecteur autour de la représentation virtuelle de la bibliothèque et de son contenu.

LES DONNÉES, UNE RÉVOLUTION ÉPISTÉMOLOGIQUE POUR LES BIBLIOTHÈQUES ?

Dans leur ouvrage intitulé *Big Data*[20], Viktor Mayer-Schoenberger et Kenneth Cukier introduisent le sujet de leur livre par ces mots :

> « Le phénomène des mégadonnées désigne tout ce qui peut être fait à une large échelle et non à une échelle plus réduite, afin d'extraire de nouvelles connaissances ou de créer de nouvelles formes de valeur, bouleversant ainsi les marchés, les organismes, les relations entre citoyens et gouvernements, et bien plus »[21].

Parce qu'elles peuvent être vues comme des institutions massivement productrices de données, il semble difficile d'envisager que les bibliothèques ne soient pas elles-mêmes touchées, d'une manière ou d'une autre, par cette (r)évolution. La question est donc de savoir quel peut être l'apport véritable du Big Data, notamment pour la production de connaissances portant sur les bibliothèques. L'existence de la section de fouille exploratoire de données (data mining) de l'OCLC montre que le monde de la documentation commence effectivement à s'intéresser à ces nouvelles méthodes. Reste à savoir dans quelle mesure ces innovations seraient réellement révolutionnaires : serait-ce parce qu'elles confèreraient davantage d'objectivité, comme le prétendent certains auteurs, ou plutôt parce qu'elles permettent, par le biais de la visualisation de données, de révéler des aspects des bibliothèques qui, jusqu'à présent, étaient restés ignorés ?

LES DONNÉES PARLENT-ELLES D'ELLES-MÊMES [22]?

Parce que les mégadonnées et leurs outils d'analyse permettraient de ne moins avoir recours à l'échantillonnage de données, porteur de marges d'erreur plus ou moins grandes, ni aux questionnaires des méthodes d'enquêtes, critiqués pour leur biais plus ou moins assumés, les acteurs du Big Data revendiqueraient de leur côté une plus grande exactitude ainsi qu'une plus grande objectivité, conférées supposément par le caractère scientifique et technique des méthodes employées. Ces méthodes reposent cependant sur des algorithmes qui traduisent en langage mathématique et mécanique les a priori de leur concepteurs. Les mégadonnées portent donc en elles une subjectivité qui est d'autant plus trompeuse qu'elle est parfois opaque et, dans certain cas, non assumée.

Des études de publics aux acteurs du Big Data

Désireuses d'améliorer les services qu'elles offrent à leurs usagers, les bibliothèques sont parfois amenées à produire des études statistiques sur leurs publics. Ce fut le cas en 2005 de la BIUSJ qui, dans le cadre de « changements fondamentaux touchant son organisation comme ses espaces », cherchait à « dresser un tableau fidèle des besoins en matière de bibliothèque au niveau de

[20] MAYER-SCHOENBERGER, Viktor et CUKIER, Kenneth, 2014.

[21] *Ibid*. « (…) big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more ». p. 6.

[22] *Ibid*. Nous faisons référence ici à un des titres de l'ouvrage de Mayer-Schoenberger et Cukier : « Letting the data speak ». p. 6.

l'université pour juger de la qualité et de l'utilité du réseau documentaire existant »[23].

La BIUSJ, et plus largement la BUPMC, a donc eu recours à une enquête quantitative par questionnaire, « outil favori, voir fétiche, dans le paysage des études sur les publics »[24], dont l'objectif est « de mieux connaître les profils sociodémographiques des publics présents (inscrits ou non inscrits), de connaître les raisons de leur présence sur les lieux, et de déterminer si leurs besoins sont satisfaits ou non »[25]. Le succès de ce type d'enquête nous invite à en examiner les caractéristiques principales, notamment le principe de l'échantillonnage : dans la mesure où les enquêtes quantitatives ont un coût et que leur cible peut être vaste, il n'est pas possible d'interroger la totalité de la population ciblée par l'enquête. On choisit donc de sélectionner un échantillon représentatif de cette population, sa représentativité devant être garantie par un tirage aléatoire et par un nombre relativement élevé de personnes interrogées, ou bien par la méthode des quotas dont le principe est de s'appuyer sur des catégories sociodémographique déterminées au préalable (grâce aux données issues des recensement de l'INSEE, par exemple) lorsque l'on souhaite reproduire la structure d'une population connue dans l'échantillon.

C'est là un point fondamental sur lequel les méthodes statistiques traditionnelles diffèrent des méthodes de traitement des données du Big Data. Deux éléments caractérisent en effet les mégadonnées : d'une part, la faculté de traiter, stocker et d'analyser des masses de données mesurées en téraoctets, et non plus de se cantonner à des échantillons limités, et d'autre part le coût moindre de la collecte de ces données, notamment parce qu'elles sont générées automatiquement, par exemple à chaque interaction d'un usager avec un service en ligne.

Dans ce contexte, la méthode de l'échantillonnage, destinée à produire un substitut à la population ciblée, deviendrait dans une certaine mesure caduque. À quoi bon, si l'on suivait donc les enseignements de Cukier et Mayer-Schoenberger, se contenter d'une petite partie des usagers d'une bibliothèque, quand on peut les avoir tous en fouillant les données produites, par exemple, par l'interaction du public avec le site internet de la bibliothèque ? Que penser également des calculs des marges d'erreurs intrinsèquement liés à cette méthode ? Considérant, par exemple, que les pourcentages calculés sur un échantillon de 1000 personnes peuvent comporter une marge d'erreur de 1,4 à 3,2 points, cette marge d'erreur serait en conséquence inexistante pour des pourcentages calculés à partir de la totalité d'une population d'utilisateurs[26].

Qui plus est, les mégadonnées apporteraient une plus grande souplesse dans leur utilisation que les échantillons de données. Dans une enquête statistique traditionnelle, en effet, on peut être amené à vouloir « distinguer plus finement les différences de réponse au sein de groupes particuliers. » Dès lors, « on pourra créer des sous-populations sur la base de certains critères qui peuvent être combinés par

[23] EVANS, Christophe (dir). *Mener l'enquête : guide des études de publics en bibliothèque*. 2011. Collection La boîte à outils. p. 130-131.

[24] *Ibid*. p. 62.

[25] *Ibid*. p. 45.

[26] Autant affirmer dès ce stade de notre réflexion que cette idée selon laquelle la science des données pourrait se passer des statistiques inférentielles est fausse, comme l'écrivent notamment O'Neil et Schutt : « In the current popular discussion of Big Data, the focus on enterprise solutions such as Hadoop to handle engineering and computational challenges caused by too much data overlooks sampling as a legitimate solution. At Google, for example, software engineers, data scientists, and statisticians sample all the time ». Non paginé dans sa version électronique. Par ailleurs, le fait de disposer de toutes les données ne signifie pas que les biais s'en trouvent effacés : « Even if we have access to all of Facebook's or Google's or Twitter's data corpus, any inferences we make from that data should not be extended to draw conclusions about humans beyond those sets of users, or even those users for any particular day ». Non paginé dans sa version électronique.

les opérateurs booléens (et, ou, sauf...) »[27]. Cependant, procéder ainsi comporte le risque d'augmenter les marges d'erreur inhérentes aux calculs qui pourraient être effectués sur les sous-populations, étant donné que l'échantillon aura été divisé. Lorsque les données sont massives, la division en sous-catégories d'études ne poserait pas ce problème, puisque les divisions s'effectueraient sur la totalité des données.

Le fait que l'on ait pris la précaution de publier des guides à propos des études des publics en bibliothèque témoigne de la longue préparation nécessaire, en amont de sa réalisation, à la méthode de l'échantillonnage. Dès lors, un échantillon peut difficilement répondre à des questions qui n'avaient pas été envisagées avant sa réalisation. Par contraste, les techniques d'analyses propres au Big Data offriraient une plus grande liberté dans les objectifs que se fixent une enquête. Leur dimension aléatoire rendrait possible la production de connaissances sans savoir au préalable ce que l'on cherche, ni quel genre d'échantillon il faudrait fournir. Alors que les échantillons ne permettent que difficilement une analyse exploratoire, être en possession de toutes (ou presque[28]) les données conférerait davantage de liberté pour les explorer, les observer sous des angles différents ou encore approfondir certains de leurs aspects.

C'est ainsi en partant du principe bancal qu'à l'ère du Big Data, nous serions en possession de toutes les données (N = tout), que Mayer-Schoenberger et Cukier prétendent que les données parleraient d'elles-mêmes.

La prétention à l'objectivité

« Un des domaines les plus significativement touché par N=tout sont les sciences sociales », écrivent Mayer-schoenberger et Cukier :

> « Elles ont perdu leur monopole sur l'interprétation des données sociales empiriques, étant donné que l'analyse des masses de données remplace les enquêteurs experts du passé. Mais lorsque les données sont collectées passivement chaque fois qu'une personne fait ce qu'elle ferait naturellement de toute façon, les anciens biais inhérents à l'échantillonnage et aux questionnaires disparaissent »[29].

C'est ici que nous aimerions discuter ce discours : « même si nous sondons absolument toutes les personnes qui quittent les bureaux de vote », écrivent O'Neil et Schutt, « nous ne comptons toujours pas les personnes qui dès le départ, ont décidé de ne pas voter. Et ces personnes pourraient bien être les personnes que nous aurions besoin de sonder afin de comprendre les problèmes de notre pays concernant le vote »[30]. En effet, les données, si massives qu'elles puissent être, ne disent pas tout : celles d'un SIGB, par exemple, n'apportent aucune information sur les personnes qui n'utilisent pas les services d'une bibliothèque et, à cet égard, les enquêtes sur les pratiques culturelles des français

[27] *Ibid*. p. 78.

[28] O'Neil et Schutt font de ce présupposé selon lequel nous disposons de toutes les données « le plus gros problème à l'ère du Big Data ». Pour étayer cela, elles prennent l'exemple de comptage des votes lors d'une élection : « Indeed, we'd argue that the assumption we make that N = all is one of the biggest problems we face in the age of Big Data. It is, above all, a way of excluding the voices of people who don't have the time, energy, or access to cast their vote in all sorts of informal, possibly unnannounced, elections. Those people, busy working two jobs and spending time waiting for buses, become invisible when we tally up the votes without them. To you this might just mean that the recommendations you receive on Netflix don't seem very good because most of the people who bother to rate things on Netflix are young and might have different tastes than you, which skews the recommendation engine toward them. But there are plenty much more insidious consequences stemming from this basic idea ». O'NEIL, SCHUTT. 2013. Non paginé dans sa version électronique.

[29] MAYER-SCHOENBERGER, CUKIER. 2013. « One of the areas that is being most dramatically shaken up by N=all is the social sciences. They have lost their monopoly on making sense of empirial social data, as big-data analysis replaces the highly skilled survey specialists of the past. The social science disciplines largely relied on sampling studies and questionnaires. But when the data is collected passively while people do what they normally do anyway, the old biases associated with sampling and questionnaires disappear ». p. 30.

[30] O'NEIL, SCHUTT. 2013. « (…) even if we poll absolutely everyone who leaves the polling stations, we still don't count people who decided not to vote in the first place. And those might be the very people we'd need to talk to to understand our country's voting problems ». Non paginé dans sa version électronique.

se révèlent toujours aussi précieuses. Dès lors, il nous paraît présomptueux de la part des auteurs d'affirmer que les méthodes propres à la science des données pourraient remplacer de manière avantageuse les statistiques traditionnelles.

Mais, au-delà de cela, que penser de l'affirmation selon laquelle les mégadonnées ferait disparaître les présupposés qu'impliquaient tout échantillon et questionnaire ? Avant toute chose, nous pourrions commencer par examiner ces présupposés : l'échantillon, comme on l'a vu, est construit en fonction d'une problématique particulière qui détermine sa constitution en ciblant une population. L'exemple le plus révélateur en est la méthode des quotas, que l'on utilise lorsque la population ciblée est connue. Quant au hasard, supposé garantir la représentativité de l'échantillon, il est difficile à obtenir de manière absolue. D'après le *Guide des études de publics en bibliothèque*, l'enquêteur doit en effet s'assurer « que le sondage respecte bien la réalité du terrain (les jours et heures d'ouverture, les lieux à prospecter dans le cadre d'un réseau, etc.) » et limiter « l'effet de proximité qui tend à s'instaurer entre l'enquêteur et l'interrogé(e), le premier ayant "naturellement" tendance à choisir des personnes qui lui paraîtront les plus abordables en fonction de leur âge, de leur sexe, de leur milieu social et aussi de leur valeurs »[31].

En ce qui concerne les questionnaires, les biais proviennent à la fois de l'enquêteur et de l'interrogé. En effet, la formulation des questions, d'une part, peut avoir pour effet d'orienter les réponses des interrogés et c'est bien en vertu de ce principe qu'il est recommandé, par exemple, de « mesurer la satisfaction au moyen d'échelles évitant le refuge vers une position moyenne »[32]. Les réponses des usagers, d'autre part, peuvent être affectées par leur propres préjugés. Ces biais sont bien connus des enquêteurs, et sont toujours pris en compte dans l'analyse des résultats. Ils auraient tendance à disparaître dans un environnement de type Big Data, à ce qu'affirme, par exemple, Andrew Nagy à propos des données générées par Summon, sorte de Google Scholar spécifiquement destiné aux bibliothèques universitaires américaines[33] :

> « Toutes les requêtes des usagers dans un même index unifié, quelque soit le degré de personnalisation de leur interface Summon locale, peuvent être vus comme la clé pour obtenir des données significatives et interprétables. Ces données peuvent mettre en évidence des comportements qui illustrent les véritables usages des services des bibliothèques, contrairement aux usages d'un petit nombre de participants observés dans des situations peu habituelles telles que celles offertes par les tâches non-ordinaire imposées à l'occasion des études d'utilisabilité »[34].

Outre la prétendue disparition de ces biais, les outils du Big Data feraient également disparaître toute hypothèse ou théorie préalable à une quelconque

[31] EVANS. 2011. p. 63.

[32] *Ibid*. p. 74.

[33] What is Summon? | University Libraries | Virginia Tech, [sans date]. [en ligne]. [Consulté le 2 août 2014]. Disponible à l'adresse : http://www.lib.vt.edu/help/summon/what-is-summon.html

[34] Data Mining « Big Data »: A Strategy for Improving Library Discovery | Blog | Serials Solutions, [sans date]. [en ligne]. [Consulté le 9 mai 2014]. Disponible à l'adresse : http://www.serialssolutions.com/en/words/detail/data-mining-big-data-a-strategy-for-improving-library-discovery. « For the past decade or longer, usability testing has been the traditional process for evaluating a software application's user experience. In usability testing, users – existing users of the application or participants recruited « off the street » – are observed while completing a series of scenarios that mimic real life examples. (…) For many years this approach has provided valuable information. However, non matter how unobtrusive the observation mechanism, users act differently when they know they are being observed. (…) All users searching across the same unified index, no matter how customized their local Summon site might be, is the key to capturing meaningful and interpretable data. This data can expose behaviors that illustrate true usage of library services, as opposed to the usage of a small number of participants being observed in an unfamiliar situation such a usability study defined tasks ».

recherche scientifique. C'est là la teneur du propos de Chris Anderson, rédacteur en chef du magazine *Wired*, prophétisant en 2008 la « fin de la théorie »[35]. Selon lui, le déluge de données rendrait la méthode scientifique obsolète, les hypothèses testées sur des données étant remplacées par des analyses reposant sur de simples corrélations, dépourvues de pré-requis. Si dans un premier temps, Mayer-Schoenberger et Cukier s'attachent à nuancer cette idée, la suite de leur propos contribuent pourtant à appuyer les propos d'Anderson :

> « À l'ère du Big Data, il n'est plus efficace de décider quelle variables examiner en s'appuyant seulement sur des hypothèses. Les jeux de données sont beaucoup trop larges et le domaine considéré probablement bien trop complexe. Heureusement, un grand nombre de contraintes qui nous poussait à une approche conduite par hypothèse ne pèse plus autant qu'auparavant. Nous avons désormais tant de données disponibles et tant de capacité de calcul que nous n'éprouvons plus le besoin de choisir laborieusement une ou plusieurs variable d'approximation et de les examiner une par une. Des analyses computationnelles sophistiquées permettent désormais d'identifier l'approximation optimale – comme cela s'est passé pour Google Flu Trends, après avoir examiné près d'un demi million de modèles mathématiques »[36].

En réalité, nous dirions plutôt que les présupposés inhérents au choix des variables, fonction des hypothèses de départ d'une recherche, se sont déplacés de ce choix des variables au choix des principes sur lesquels reposent les algorithmes (les « analyses computationnelles complexes ») permettant éventuellement de choisir ces variables à notre place[37].

De fait, lorsque les auteurs faisant la promotion du Big Data pour son objectivité, opposent une conception scientifique traditionnelle aux nouvelles méthodes d'analyse du Big Data, ils opposent implicitement une conception de la science moderne et platonicienne à une conception antique et aristotélicienne : en parlant d'elles-mêmes, les données massives remettraient Aristote au goût du jour. Mais les données parlent-elles véritablement d'elles-mêmes ? Rien n'est moins sûr, car ce serait oublier que les techniques d'analyse employées ont largement recours aux algorithmes, et de ce fait, aux mathématiques. Or, « pour retrouver Aristote, écrit Olivier Rey, il faudrait oublier non telle ou telle théorie, mais le cadre mathématique lui-même – ce qui n'est plus en notre pouvoir lorsqu'il s'agit d'interroger scientifiquement la nature »[38]. Il est donc erroné de penser que les données puissent parler d'elles-mêmes : « Est-ce que vraiment N = tout ? » se demandent O'Neil et Schutt. « C'est bien là le problème : ce n'est quasiment jamais tout. Et nous passons souvent à côté de ce à quoi nous devrions prêter le plus attention[39] ». Par la seule prétention selon laquelle une variable suffirait à représenter un phénomène complexe, les algorithmes continuent à faire revivre le préjugé galiléen qui consiste à voir en l'univers un livre écrit mathématiquement. Et en ce qui concerne

[35] « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete ». *WIRED*. Consulté le 2 août 2014. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

[36] MAYER-SCHÖNBERGER, CUKIER. 2013. p. 55. « In the big-data age, it is no longer efficient to make decisions about what variables to examine by relying on hypotheses alone. The data sets are far too big and the area under consideration is probably far too complex. Fortunately, many of the limitations that forced us into a hypothesis driven approach non longer exist to the same extent. We now have so much data available and so much computing power that we don't have to laboriously pick one proxy or a small handful of them and examine them one by one. Sophisticated computational analysis can now indentify the optimal proxy – as it did for Google Flu Trends, after plowing through almost half a billion mathematical models ».

[37] *Cf* O'Neil et Schutt à propos de l'extraction de variable (feature extraction) : « This process we just went through of brainstorming a list of features for Chasing Dragons is the process of *feature generation* or *feature extraction*. This process is as much of an art as a science. It's good to have a domain expert around for this process, but it's also good to use your imagination ». O'NEIL, SCHUTT. 2013. Non paginé dans sa version électronique. On voit par là que la sélection de variables, même au moyen d'algorithme, reste un processus subjectif.

[38] MAYER-SCHÖNBERGER, CUKIER. 2013. p. 55. et p. 60.

[39] O'NEIL, SCHUTT. 2013. « Can N = all ? Here's the thing : it's pretty much never all. And we are very often missing the very things we should care about most ». Non paginé dans sa version électronique.

l'application des algorithmes aux données sociales, les présupposés vont plus loin encore que ceux impliqués simplement par la théorie et les mathématiques, comme l'écrit Ronald E. Day à propos des analyses bibliométriques :

> « Ce que nous découvrons, c'est que les algorithmes d'informatique sociale, telle que PageRank (algorithmes d'analyse des liens) et les systèmes de recommandation, renforcent la « véracité » des lois bibliométriques (telles que la loi de Lotka), simplement parce qu'ils automatisent les théories comportementales inhérentes à de telles « lois » puis réinfusent cela dans le comportement des utilisateurs. (…) Les découvertes empiriques ne parlent jamais d'elles-mêmes. (...) Dès lors, les « objets » d'étude et leur mesures empiriques (ainsi que les outils et algorithmes qui y participent) ne seraient rien d'autre que des instruments de réaffirmation de normes sociales, culturelles et politiques. La seule chose qu'ils affirment est la certitude de l'idéologie »[40].

Les algorithmes contribuent donc à transposer les présupposés idéologiques de leur concepteurs, et à les ré-infuser dans nos comportements lorsque ces algorithmes sont utilisés non plus seulement pour analyser des données sociales mais pour les générer. Ces présupposés, quels sont-ils et comment les mettre à jour ?

Les algorithmes au regard critique de la sociologie

L'exemple de Google, désigné comme la parangon des compagnies s'appuyant sur les méthodes du Big Data, nous paraît emblématique de ce phénomène politique qui consiste à revendiquer une objectivité dans le traitement des données en arguant du fait que les procédés utilisés sont technologiques et non humains. En réaction à cela, Tarleton Gillespie écrit dans son essai sur la pertinence des algorithmes :

> « Ce dont nous avons besoin, c'est d'une interrogation des algorithmes en tant que caractéristique clé de notre écosystème informationnel, et des formes culturelles émergeant dans leurs ombres, avec une attention particulière portée à l'endroit et à la manière avec laquelle l'introduction d'algorithmes dans nos pratiques de connaissance humaine peuvent avoir des ramifications politiques »[41].

Dans un article intitulé « La subjectivité algorithmique et le besoin d'être informé »[42], Neal Thomas semble avoir répondu à Tarleton Gillespie. Son analyse

---

[40] DAY, Ronald E. « "The Data – It is *Me* !" ("Les données – c'est *Moi !*") » dans CRONIN, Blaise et SUGIMOTO, Cassidy R., 2014. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Cambridge, Massachusetts : MIT Press. p.70-71. « What we find is that social computing algorithms, such as PageRank (link analysis algorithms) and recommender systems, strengthen the « truthfulness » of bibliometric « laws » (such as Lotka's law), simply because they automate the group behavioral assumptions inherent in such « laws » and then feed this back into user behavior. (…) Empirical findings never simply show themselves. Citation analytics, either explicitly or implicitly, as a social science *must* indicate social explanations of various types of regular behaviors. Once again, the epistemic problem of social science operationalization – which becomes political and psychological when citation analyses are highly valued in restricted (e.g. Academic) or general (e.g. Social) economies – is what happens when these explanations are the very basis for the metrics to begin with. Then, the « objects » of study and their empirical measurements (and the tools and algorithms that aid this) may be nothing other than devices in the restaing of social, cultural, and political norms. What they would assert is the certainty of ideology ».

[41] GILLESPIE, Tarleton. « The relevance of algorithms », à paraître dans Gillespie, Tarleton, BOCZCOWSKI, Pablo et KIRSTEN, Foot. *Media Technologies*. Cambridge, MA : MIT Press. Consulté le 3 août 2014 à l'adresse Web : http://www.tarletongillespie.org/essays/Gillespie%20-%20The%20Relevance%20of%20Algorithms.pdf. « What we need is an interrogation of algorithms as a key feature of our information ecosystem (…), and of the cultural forms emerging in their shadows (…), with a close attention to where an din what ways the introduction of algorithms into human knowlegde practices may have political ramifications ».

[42] THOMAS, Neal. 2012. « Algorithmic subjectivity and the need to be in-formed. » dans LATZKO-TOTH, Guillaume, MILLERAND, Florence. *TEM 2012 : Proceedings of the Technology & Emerging Media Track – Annual*

des algorithmes permet de mettre en évidence les présupposés socioprofessionnels qui ont présidé à l'évolution des algorithmes utilisés pour la recherche documentaire. De son point de vue en effet, l'algorithme renvoie à cette capacité humaine à transposer sous la forme logique du langage informatique ce qui relève du signe, à savoir l'expression formelle d'un objet ou d'un concept. Google, par exemple, transpose en langage informatique et logique la représentation qu'il se fait du besoin dans le processus de recherche d'information. C'est en ce sens que l'on peut dire que Google est un « médium algorithmique »[43] : l'algorithme est le moyen par lequel « la pensée est littéralement faite mécanique »[44].

Ces représentations théorisées du processus de recherche d'information ne deviennent évidentes que lorsqu'on retrace l'évolution des algorithmes qui les ont modélisés, ce qu'a fait Neal Thomas. Si donc l'on en croit son propos, l'informatique traditionnelle a d'abord défini le besoin d'information comme la simple correspondance entre le besoin d'un document spécifique et le document lui même : « Pour le dire en quelques mots, le besoin était essentiellement exprimé à travers la forme de la requête sémantiquement précise : "j'ai besoin de trouver le document spécifique dont je présume qu'il est appelé *x*" »[45].

Cette première théorisation du processus de recherche a ensuite évolué vers une autre théorie, celle du besoin cognitif, influencée cette fois non par l'informatique mais par le milieu des bibliothèques et des sciences de l'information : l'algorithme devait cette fois « modéliser l'interaction vécue entre un bibliothécaire de référence et une personne venue demander des renseignements »[46]. Cette modélisation devait prendre en compte le processus par lequel la recherche « passait par des phases d'adaptation communicative entre les acteurs, à savoir le bibliothécaire cherchant à découvrir le document qui pourrait répondre à la question-connaissance de la personne en recherche de renseignement »[47].

Là dessus, la théorie du besoin de l'information évolue encore, cette fois sous l'influence d'une conception économique de l'individu, à savoir « une correspondance utilitarienne-économique entre le sujet et l'objet. Plus concrètement, elle est basée sur la théorie du choix rationnel »[48]. Google conçoit désormais le besoin d'information comme « la formulation et la satisfaction de "situations problèmes" en cours et socialement contextualisée »[49]. Dès lors, le besoin d'information serait déterminé en grande partie par ses propres traces et par les comportements passés de précédents utilisateurs. Dans ce contexte, l'algorithme k-NN était celui qui transposait le mieux la théorie du besoin d'information propre à Google, puisqu'il « réorganise perpétuellement un "voisinage" de traces pour les utilisateurs présents en fonction des chemins tracés par les précédents »[50].

Conference of the Canadian Communication Association (Waterloo, May 30 D June 1, 2012). Consulté le 3 août 2014. http://www.tem.fl.ulaval.ca/www/wpcontent/PDF/Waterloo_2012/THOMASFTEM2012.pdf

[43] *Ibid*. p. 2.

[44] *Ibid*. « The efficiency for human beings can be found where thinking can literally be made mechanical ». p. 3.

[45] *Ibid*. « A focus on an *instrumental* need for a *specific document*, that follows a simple 'best-match' engineering principle. (…) search was simply a matter of correct encoding and deconding. To put it in a phrase, need was essentially expressed through the form of the *semantically precise query* : "I need to find the specific document I believe is called *x*" ». p. 4-5.

[46] *Ibid*. « Taylor was concerned to model the lived interactions between a reference librarian and an inquirer ». p. 6.

[47] *Ibid*. « He especially sought to account for how the inquiring process went through communicative *phase of adaptation* between the actors, the librarian seeking to discover the document that answered the knowledge-question of the inquirer ».

[48] *Ibid*. « the theoretical framework is a *utilitarian-economic* correspondence between subject and object. More simply, it is based in rational choice theory ». p. 4

[49] *Ibid*. « Contemporary network interfaces like Google rely on the collective posing and satisfaction of ongoing, socially contextualized 'problem situations' ».

[50] *Ibid*. « perpetually reorganizing a 'neighborhood' of records for present users according to paths laid down by prior ones ». p. 9. Sur les présupposés inhérents à des algorithmes comme k-NN, on peut consulter O'Neil et Schutt : « The k-NN algorithm is an example of a nonparametric approach. You had non modeling assumptions about the underlying data-generating distributions, and you weren't attempting to estimate any parameters. But you still made *some* assumptions, which were :

- Data is in some feature space where a notion of « distance » makes sense.
- Training data has been labeled or classified into two or more classes.

En évoquant cette démonstration faite par Neal Thomas de la subjectivité des algorithmes, il apparaît que ces derniers sont amenés à évoluer non pas tant en fonction d'une recherche constante d'amélioration et d'efficacité, qu'en fonction des aléas d'une concurrence entre des visions radicalement différentes de la société et de ses besoins. La conception économique du besoin d'information, peut-être du fait de la prédominance que lui confère la pensée contemporaine, l'a emporté sur les autres théories provenant de l'informatique et des bibliothèques. À cet égard, l'efficacité de Google ne se mesurerait pas tant à sa capacité à apporter des réponses pertinentes à nos questions qu'à sa capacité à transposer de manière adéquate dans ses algorithmes un cadre conceptuel dominant.

Ces algorithmes qui peuvent servir à faire parler les données sont donc eux-mêmes des média, dont les évolutions transcrivent des luttes politiques entre des visions sociales différentes. Mais malgré son caractère mythologique, la revendication de l'objectivité des algorithmes, à travers celle de leur impartialité, continue à être régulièrement mis en avant par les acteurs du Big Data :

> « Par dessus toute autre chose, les fournisseurs d'algorithmes informationnels doivent affirmer que leurs algorithmes sont impartiaux. L'effectivité de l'objectivité algorithmique est devenue fondamentale au maintien de ces outils comme courtiers de la connaissance pertinente. Aucun fournisseur n'a plus insisté sur la neutralité de ses algorithmes que Google, qui répond régulièrement aux demandes qui lui sont adressées de modifier les résultats de ses recherches par l'affirmation que l'algorithme ne doit pas être manipulé »[51].

Les méthodes analytiques du Big Data offrent bien des avantages par rapport aux statistiques traditionnelles, notamment au regard de leur faible coût et de la souplesse qu'elles offrent. Mais là où les enquêtes de publics assumaient et prenaient en compte dans leurs résultats la subjectivité inhérente à leur élaboration, les outils du Big Data et leurs algorithmes, au contraire, peuvent parfois revendiquer leur objectivité tout en modélisant, de manière implicite, des présupposés théoriques. Dès lors, il devient nécessaire, avant toute utilisation d'un algorithme dans un projet d'analyse des données, de prendre en compte les aspects politiques[52] qui peuvent lui être attachés. Mais au-delà du simple enjeu épistémologique et de la volonté d'honnêteté intellectuelle, la prise en compte et la reconnaissance des idéologies qui façonnent notre recherche de l'information sur internet et dans les bibliothèques est aussi, on l'aura compris, un enjeu démocratique.

- You pick the number of neighors to use, **k**.
- You're assuming that the *observed features* and the *labels* are somehow associated. They may not be, but ultimately your evaluation will help you determine how good the algorithm is at labeling. You might want to add more features and check how that alters the evaluation metric. You'd then be tuning both *which* features you were using and *k*. But as always, you're in danger of overfitting ». O'NEIL, SCHUTT. 2013. Non paginé dans sa version électronique.

[51] GILLEPSIE, à paraître. « Above all else, the providers of information algorithms must assert that their algorithm is impartial. The performance of *algorithmic objectivity* has become fundamental to the maintenace of these tools as legitimate brokers of relevant knowledge. No provider has been more adamant about the neutrality of its algorithm than Google, which regularly responds to requests to alter their search results with the assertion that the algorithm must not be tampered with ».

[52] L'aspect politique des algorithmes est précisément ce qui doit être mis en avant, davantage que les aspects techniques qui les concernent : *Ibid*. « In attempting to say something of substance about the way algorithms are shifting our public discourse, we must firmly resist putting the technology in the explanatory driver's seat. While recent sociological study of the Internet has labored to undo the simplistic technological determinism that plagued earlier work, that determinism remains an alluring analytical stance. A sociological analysis must not conceive of algorithms as abstract, technical achievements, but must unpack the warm human and institutional choices that lie behind these cold mechanisms ».

Il convient maintenant de se demander dans quelle mesure l'analyse des données des bibliothèques peut prendre en compte, ou non, la subjectivité inhérente à la science des données.

L'exemple de l'Online Computer Library Center (OCLC)

L'OCLC dispose d'une section entièrement consacrée à la science des données. Cette dernière s'attache à produire des rapports sur l'évolution des collections physiques des bibliothèques américaines dans le contexte de la numérisation de masse.

Une section consacrée à l'extraction et à l'analyse de données

Organisation mondiale à but non lucratif dédiée aux bibliothèques, mais aussi organisme de recherche, l'OCLC se prêtait tout particulièrement à la réutilisation des données bibliographiques. C'est elle qui, en effet, est derrière le pilotage de WorldCat, considéré comme le plus grand catalogue OPAC du monde. De fait, l'OCLC s'est engagée depuis 2012 dans un processus d'ouverture de ses données bibliographiques en envisageant « la création d'une réserve mondiale de données partagées qui pourrait être utilisée et réutilisée pour la description des ressources, réduisant ainsi le travail redondant, inhérent aux processus actuels de catalogage »[53]. Mais la mutualisation du catalogage n'est pas la seule réutilisation envisagée par l'OCLC. L'organisme de recherche s'est en effet doté d'une section entièrement consacrée à l'extraction et à l'analyse de données (Data Mining Research Area) , les objectifs assignés à cette section étant les suivants :

> « En savoir plus sur les caractéristiques propres aux collections des bibliothèques.
> Générer des présentations intéressantes et innovantes des données.
> Fournir des informations pour répondre à un certain nombre de besoins en matière de prises de décision dans les bibliothèque, tels que :
> - le développement des collections,
> - la numérisation,
> - la conservation »[54].

Ces objectifs ont été déclinés à l'échelle de la section de recherche de l'OCLC en plusieurs projets, notamment « l'analyse de la taille et des caractéristiques des collections des fonds agrégés d'imprimés, avec une emphase sur leurs implications pour les décisions à prendre en matière de numérisation et de conservation », la déduction par inférence « des publics cibles, ou des niveaux d'audience des ouvrages à partir des informations provenant des fonds » ou encore « l'évaluation comparative de collections : l'étude du développement, de l'évaluation et du partage des ressources pour les collections imprimées et électroniques »[55], entre autres choses. On le voit, ces projets ont pour beaucoup à voir avec ce que nous considérons comme relevant de la politique documentaire, à savoir l'ensemble des décisions ayant trait à l'acquisition, la gestion et la mise en valeur des collections des bibliothèques.

Certains d'entre eux, comme celui qui concerne les niveaux d'audience, ont permis le développement de services devant être intégrés à l'interface de recherche de WorldCat. Ils ont été présentés en 2013 à Strasbourg lors du meeting du conseil régional de l'EMEA, par Roy Tenant, gestionnaire principal de projet de l'OCLC[56]. Roy Tenant insiste particulièrement sur l'élaboration des « identités WorldCat » qu'il décrit comme

[53] CARTIER, Aurore, 2012. *Bibliothèque et Open data. Et si on ouvrait les bibliothèques sur l'avenir ?* Consulté le 15 décembre 2014. Disponible à l'adresse Web : http://www.enssib.fr/bibliotheque-numerique/documents/60401-bibliotheque-et-open-data-et-si-on-ouvrait-les-bibliotheques-sur-l-avenir.pdf. p. 61.

[54] ADMIN, 2012. Data Mining Research Area. [en ligne]. 4 août 2012. [Consulté le 29 janvier 2014]. Disponible à l'adresse : http://oclc.org/research/activities/mining.html

[55] *Ibid*.

« algorithmiquement construite à partir de la base de données de WorldCat »[57]. Le principe des identités est en effet de rassembler sur une même page l'ensemble des données concernant un auteur ou créateur, en les extrayant de la base de données à l'aide d'algorithmes et de programmes.

En somme, l'idée est de rassembler toutes les informations possibles sur un auteur à partir de données dispersées. Ces informations sont de plusieurs types. Citons dans l'ordre : la période de publication de l'auteur, ainsi que de l'ensemble des œuvres qui ont été publiées sur lui, l'ensemble des formes sous lesquelles le nom de cet auteur se rencontre, l'ensemble des langues dans lesquelles il a été publié, les œuvres de cet auteur les plus possédées par les bibliothèques, le niveau du public visé par ses œuvres (jeunesse, général ou spécialisé), des liens vers le fichier VIAF de cet auteur, mais aussi vers l'article Wikipédia qui le concerne (inversement, des liens ont été inclus dans les articles de Wikipédia pointant vers le fichier VIAF) et, enfin, le nuage de sujets couverts par cet auteur, permettant de voir quel thèmes principaux sont associés à son œuvre. Toutes ces informations doivent permettre d'enrichir la navigation de l'utilisateur sur l'interface de WorldCat, ces enrichissements pouvant être vus comme la valeur ajoutée apportée par l'application de la science des données aux données bibliographiques.

Si donc la section de « data mining » de l'OCLC a recours à des outils relevant de la science des données, quels peuvent être les présupposés qui leur sont inhérents ?

L'algorithme « Work-Set FRBR »

En ce qui concerne l'élaboration des identités WorldCat, il ne nous a guère été possible de trouver des informations sur les algorithmes qui ont été utilisés pour les effectuer. En revanche, nous avons pu trouver sur le site de l'OCLC de la documentation concernant un algorithme « Work-set FRBR » (traduisons par « groupe-œuvre FRBR », FRBR désignant les Functional Requirement for Bibliographic Records) qui consiste pour sa part à rassembler toutes les informations concernant non pas un auteur ou un créateur mais une œuvre. Il s'agit donc du même principe que les identités WorldCat, mais appliqué aux œuvres, telles qu'elles sont définies par les FRBR.

Pour comprendre le fonctionnement de l'algorithme, peut-être est-il bon de rappeler comment fonctionne les FRBR : les Spécifications Fonctionnelles des Notices Bibliographiques sont un modèle conceptuel de notices bibliographiques dont l'objectif est de fournir un cadre commun à la rédaction de ces notices. D'après Wikipédia, « elles sont conçues comme un outil pour l'établissement de futures normes bibliographiques ».[58] Plus concrètement, les FRBR distinguent quatre mentions « essentielles » devant être identifiables dans toutes les notices : tout d'abord l'« œuvre », produit intellectuel d'un auteur ou d'un créateur, puis son « expression », qui peut être toute réalisation créée à partir de cette œuvre, telle qu'une traduction. La « manifestation » doit ensuite représenter la matérialisation de cette expression, telle que l'édition particulière d'une traduction. Enfin, le « document » représente l'exemplaire, tel que celui de l'édition de la traduction d'une œuvre.

[56] *Leveraging WorldCat: Data Mining the largest library database in the World*, 2013. [en ligne]. [Consulté le 14 juillet 2014]. Disponible à l'adresse : http://www.youtube.com/watch?v=atA2QadzTdY&feature=youtube_gdata_player

[57] *Ibid*.

[58] Spécifications fonctionnelles des notices bibliographiques, 2014. *Wikipédia* [en ligne]. [Consulté le 4 août 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Sp%C3%A9cifications_fonctionnelles_des_notices_bibliographiques&oldid=103576162.

C'est précisément ce cadre conceptuel qui va être repris dans le but de définir un algorithme permettant de regrouper ensemble des notices qui ont trait à la même œuvre, et ce toujours dans le but de faciliter et d'enrichir la navigation des utilisateurs de WorldCat : les FRBR permettent en effet de définir des critères selon lesquelles on pourra classer ensemble des notices qui se ressemblent. Les ensembles de ces notices rassemblées selon ces critères sont appelés « groupes-œuvre ». Pour constituer ces groupes, l'algorithme devra attribuer à chacune des notices une clé unique, sur la base desquelles ces notices seront regroupées ensemble, comme l'explique Thomas Hickey :

> « Le but est de créer une clé capable d'identifier de manière sûre et unique un groupe-FRBR. Le cas le plus aisé est celui où nous avons un auteur et un titre, ou un titre solitaire et uniforme. Si nous n'avons pas un auteur ou un titre uniforme, alors nous essayons de trouver des champs correspondants au nom (les étiquettes 7XX) pour aider à identifier des documents associés. Les notices qui ne possèdent que des champs 24X (il n'y pas de champ 1XX ou 7XX dans la notice) sont combinées avec leur nombre WorldCat pour construire une clé unique. Nous ne pouvons pas combiner ces titres qui s'associent, étant donné que nous n'avons pas assez d'informations pour grouper de manière fiable ces documents »[59].

Les contournement développés pour palier au fait qu'une œuvre peut être sans titre ni auteur montrent bien que le concept d'œuvre tel qu'il est défini par les FRBR et qui commande l'algorithme que nous venons de décrire n'a rien d'évident. Pour le démontrer, David Weinberger prend l'exemple d'*Hamlet*[60] : si l'on suit en effet la description FRBR, *Hamlet* constitue bien une œuvre (au sens platonicien du terme, puisqu'elle n'a jamais existé en tant que telle), « de par toutes les manières différentes avec lesquelles elle a été jouée et publiée »[61]. La version d'*Hamlet* incluse dans le Premier Folio constitue alors une des expression de l'œuvre, les impressions ou enregistrements qui en ont été faits, ses manifestations. Chaque exemplaire de ses manifestation, on l'a vu, constitue alors un document de l'œuvre *Hamlet*. « Tout cela semble assez clair, écrit David Weinberger, mais cela se complique rapidement ».

> « La version d'*Hamlet* réécrite pour les enfants avec une fin heureuse est-elle encore *Hamlet* ? Et que penser des œuvres inspirées par *Hamlet*, telles que le *Rosencrantz et Guildenstern sont morts* de Tom Stoppard et le *Sortir avec Hamlet : L'histoire d'Ophélie* de Lisa Fiedler ? Les FRBR disent que lorsque la modification d'une œuvre "implique un degré signifiant de travail artistique ou intellectuel indépendant", elle devient une nouvelle œuvre »[62].

L'algorithme des « ensembles-FRBR » pose donc la question fondamentale de l'identité de l'œuvre. Dans sa nouvelle intitulée « Pierre Ménard, auteur du *Quichotte* »[63], Borges exposait les termes du problème : Ménard n'essaye pas d'écrire un nouveau *Quichotte*, mais *le Quichotte*. Son texte est identique à celui de Cervantès,

---

[59] HICKEY, Thomas B., TOVES, Jenny. 2009. « FRBR Work-Set Algorithm, v. 2.0 ». OH: OCLC Online Computer Library Center, Inc. (Research division). Consulté le 4 août 2014 à l'adresse Web : http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf. « The goal is to create a key that can uniquely and confidently identify a work-set. The best cases occur when we have an author with a title or a solitary uniform title. If we don't have an author or a uniform title then we try to find name fields (7XX tags) to help identify related items. Records that only have a 24X field (no 1XX or 7XX fields exist in the record) get combined with their Worldcat number to force unique keys. We cannot combine those matching titles since we don't have enough information to reliably group the items ».

[60] WEINBERGER, David. 2008. *Everything Is Miscellaneous: The Power of the New Digital Disorder*. Henry Holt and Company.

[61] *Ibid*. « The most abstract concept of [FRBR] describes is a *work*, such as *Hamlet* in all the different ways it is performed and published ». p.251.

[62] *Ibid*. « All this sound quite neat, but it gets messy quickly. Is the version of *Hamlet* rewritten for children with a happy ending still *Hamlet* ? How about works inspired by *Hamlet*, such as Tom Stoppard's *Rosencrantz et Guildenstern Are Dead* and Lisa Fiedler's *Dating Hamlet : Ophelia's Story* ? The FRBR says that when the modification of a work "involves a significant degree of independent intellectual or artistic effort," it becomes a new work ».

[63] BORGES, Jorge Luis. 1944. « Pierre Ménard, auteur du *Quichotte* » dans *Fictions*. Éditions Gallimard.

mais il ne le plagie pas et Borges soutient qu'il s'agit d'une autre œuvre puisque produite à trois siècle d'intervalle et, par là, écrite dans une autre perspective, selon d'autres valeurs que celles qui avaient présidé à l'écriture du *Quichotte* de Cervantès. Les termes du débat sont donc les suivants[64] : doit-on identifier l'œuvre du point de vue de sa conception, de son autorité, comme le fait l'algorithme de l'OCLC, ou bien de sa réception et de son interprétation, comme le ferait l'algorithme d'Amazon qui effectue ses regroupements en fonction des goûts passé des usagers, exprimés par leurs activités de téléchargement, lecture, commentaires, notations, ou par le fait de mettre le document dans ses favoris, par exemple[65] ? Si l'algorithme FRBR de l'OCLC échoue à regrouper l'intégralité des œuvres en fonction de leur autorité, puisque nous avons vu que cela était loin d'être une évidence, l'algorithme d'Amazon y réussit-il davantage en cherchant à déterminer la réception de l'œuvre ? Pour Tarleton Gillespie, la réponse est négative :

> « Dans ces cycles d'anticipation, ce sont les bits d'information qui sont les plus lisibles pour l'algorithme et qui, ainsi, ont tendance à représenter les utilisateurs. Facebook sait beaucoup de ses utilisateurs, mais cependant, il ne sait que ce qu'il est capable de savoir. L'information la plus connaissable (la géo-localisation, la plate-forme informatique, les informations du profile, les amis, les mises à jour de statut, les liens suivis sur un site, le temps passé sur un site, l'activité sur un autre site comportant des boutons "j'aime" ou des cookies) est un rendu de l'usager, un "dossier numérique" (…) ou une "identité algorithmique" (…) qui est imparfaite mais suffisante[66]. Ce qui est moins lisible ou ne peut pas être connu des utilisateurs est tombé dans l'oubli ou est grossièrement approché. Comme Balka (2011) l'écrivait, les systèmes d'information produisent des "corps-ombres" en mettant l'accent sur certains aspects de leurs sujets et en passant sur d'autres »[67].

Là où nous aurions tendance à nous offusquer de l'immixtion de Google dans nos vies privées, c'est au contraire sa volonté tenace de nous classer en dépit de son insuffisance à nous cerner qu'il faudrait dénoncer. C'est d'ailleurs la raison pour laquelle nous n'avons pas souhaité nous appesantir dans cette étude sur les enjeux de vie privée propres au Big Data.

Mais si l'on peut dire qu'un algorithme utilisé par l'OCLC pour naviguer dans WorldCat est d'une certaine manière biaisé, que penser alors des recherches plus globales sur les bibliothèques américaines dans lesquels il a été utilisé ?

---

[64] On retrouve le même débat, certes formulé autrement, dans O'Neil et Schutt : « We could do a Google search for "data science" and perform a text-mining model. But that would depend on us being a *usagist* rather than a *prescriptionist* with respect to language. A usagist would let the masses define data science (where "the masses" refers to whatever Google's search engine finds). Would it be better to be a prescriptionist and refer to an authority such as the *Oxford English Dictionary* ? Unfortunately, the *OED* probably doesn't have an entry yet, and we don't have time to wait for it. Let's agree, that there's a spectrum, that one authority doesn't feel right and that "the masses" doesn't either ». O'NEIL, SCHUTT. 2013. Non paginé dans sa version électronique.

[65] THOMAS. 2012.

[66] Il nous semble que cette volonté de ce contenter de profils numériques et algorithmiques pour prendre des décisions concernant notre individualité est précisément le problème que soulignent également Antoinette Rouvroy et Thomas Bern. ROUVROY, Antoinette et BERNS, Thomas, 2013. Gouvernementalité algorithmique et perspectives d'émancipation. *Réseaux*. 1 avril 2013. Vol. 177, n° 1, pp. 163-196. « Sans considérer ceci comme vain, nous voulons signaler ici avec force l'indifférence de ce « gouvernement algorithmique » pour les individus, dès lors qu'il se contente de s'intéresser et de contrôler notre « double statistique », c'est-à-dire des croisements de corrélations, produits de manière automatisée, et sur la base de quantités massives de données, elles-mêmes constituées ou récoltées « par défaut ». Bref, ce que nous sommes « en gros », pour reprendre la citation d'Éric Schmidt, ce n'est justement plus aucunement nous-mêmes (êtres singuliers). Et c'est justement cela le problème, problème qui, comme nous le verrons, relèverait plutôt d'une raréfaction des processus et occasions de subjectivation, d'une difficulté à devenir des sujets, que d'un phénomène de « désubjectivation » ou de mise en danger de l'individu ». p. 180.

[67] GILLESPIE, à paraître.

Une des publications de l'OCLC : « Livres sans frontières »

Un exemple de l'utilisation de l'algorithme que nous venons de décrire peut-être observé dans la publication de Brian Lavoie et Roger Schonfeld intitulée « Livres sans frontières, un bref horizon de la collection d'imprimés à l'échelle du système »[68]. Partant du principe que « de nos jours, les décisions ayant trait à un nombre important de domaine tirerait bénéfice de considération provenant du contexte plus large du système »[69], les auteurs se posent la question de savoir de quelle manière la collection globale décrite par les données de WorldCat se répartit à l'échelle des différentes institutions qui y participent et en quoi ces informations pourraient, dans un futur proche, influencer les politiques de numérisation, de conservation et de médiation des collections.

Les conclusions des analyses des auteurs effectuées sur les quelques 32 millions de données provenant des notices des livres imprimés contenues dans WorldCat sont les suivantes. Un premier constat est qu'il existe en moyenne 1,2 manifestation par œuvre[70]. S'interrogeant ensuite sur le degré de redondances des différentes collections entre elles, les auteurs font part de leur admiration devant la relativement faible part de redondance observée. Vient alors la répartition des collections par dates de publication : Lavoie et Schonfeld observent que la moitié des documents ont été publiés après 1977, ce qui témoigne donc d'une forte accélération de l'activité éditoriale dans le dernier tiers du XX^{e} siècle. Quant aux langues des publications : un peu plus de la moitié des livres dont les notices sont contenues dans WorldCat sont publiés en anglais, les autres langues majoritaires étant l'allemand et le français. Enfin, une estimation très grossière permet d'affirmer que depuis 1940, près de 48% de la littérature globale, toutes disciplines confondues, est couverte par les fonds dont les notices sont enregistrées dans WorldCat.

Que penser de ces observations ? Certes, l'utilisation de l'algorithme FRBR n'a-t-elle peut-être pas tant d'influence sur les résultats finaux des analyses, il est en tout cas difficile de le savoir. Le problème, s'il doit y en avoir un, réside peut-être d'avantage dans le fait que l'analyse des données à un niveau aussi large permet effectivement de se faire une idée des caractéristiques fondamentales des collections globales et institutionnelles, mais guère sur le pourquoi ni sur le comment de ces chiffres. En réalité ces analyses ne nous parlent pas car elles ne font pas intervenir la subjectivité des professionnels qui ont œuvré pour la constitution de cette collection globale : il manque peut-être aux rapports de l'OCLC un aller-retour entre les données et les discours des bibliothécaires qui les ont conçus, tel qu'on pourrait l'observer notamment dans l'essai *The life and death of metadata*[71].

Il est peut-être nécessaire, dans ce contexte, de se mettre à la recherche d'autres approches qui permettraient de compléter ces premières analyses. C'est dans cette perspective que nous nous proposons maintenant d'interroger l'apport de la visualisation de données dans le cadre des humanités numériques.

[68] LAVOIE, Brian F., SCHONFELD, Roger C. «Books without Boundaries : A Brief Tour of the System-wide Print Book Collection » dans DEMPSEY, Lorcan, LAVOIE, Brian F., MALPAS, Constance, CONNAWAY, Lynn S., SCHONFELD, Roger C., SHIPENGROVER J.D. et WAIBEL, Günter. 2013. *Understanding the Collective Collection : Towards a System-wide Perspective on Library Print Collections*. Dublin, Ohio : OCLC Research. Consulté le 5 août 2014. Disponible à l'adresse Web : http://oclc.org/research/publications/library/2013/2013-09.pdf.

[69] *Ibid*. p. 9.

[70] On voit ici une application de l'algorithme des « groupes-œuvre FRBR ».

[71] The Life and Death of Data, [sans date]. *op. cit*.

UNE MANIÈRE INNOVANTE DE PRODUIRE DES CONNAISSANCES SUR LES BIBLIOTHÈQUES : LA VISUALISATION DE DONNÉES.

Les humanités numériques apportent un cadre critique aux nouvelles possibilités introduites par les outils du Big Data, et permettent de renouveler la manière de produire des connaissances, notamment par le biais d'une visualisation des données prenant en compte la subjectivité de l'observateur. L'exemple de l'Observatoire de la Bibliothèque, développé par le Metalab d'Harvard, illustre ce propos.

La visualisation au regard critique des humanités numériques

Dès le début du XIX[e] siècle, l'administration française a rendu familière la visualisation de données, ou plus exactement certaines de ses composantes que sont les graphiques et les cartes. Sous Napoléon, l'administration est en effet grande consommatrice de statistiques : l'instrumentalisation de ces dernières au service de desseins politiques ont éloigné pendant longtemps la possibilité d'un regard auto-critique et scientifique sur cette discipline, comme l'écrit Johanna Drucker dans son essai intitulé *Graphesis : la production et la représentation visuelle de connaissances*[72] . L'héritage positiviste et politique qui sous-tend donc l'emploi de la visualisation données conduit ses premiers théoriciens à en revendiquer l'objectivité et l'auto-évidence : à l'instar des données qu'elle est censée représenter, la visualisation parlerait d'elle-même. C'est notamment ce qu'explique Edward Tufte, considéré comme l'un des premiers penseurs de la visualisation de données, lorsqu'il recommande à l'infographe de « montrer les données » et « d'éviter de transformer ce que les données ont à dire »[73].

Dans ce contexte, le rôle des humanités numériques consiste non seulement à élaborer un cadre critique en replaçant chacune des composantes de la visualisation de données dans le contexte théorique qui l'a vu naître[74], mais également à promouvoir un enrichissement théorique de la visualisation au moyen des textes de praticiens-théoristes du XX[e] siècle qui ont fondé l'enseignement du graphisme et du design : Wassily Kandinsky, Laszlo Moholy-Nagy et Paul Klee, pour ne citer qu'eux. « L'histoire culturelle des infographies, des diagrammes, des cartes, des graphiques et des autres images schématiques, écrit Johanna Drucker, est un champ riche à explorer pour trouver des modèles productifs à l'horizon des outils théoriques fournis par les humanités »[75].

Ainsi considérée au regard des humanités Numériques, la visualisation de donnée peut être vue comme plus informative du point de vue de ses méthodes que les autres techniques incluses dans le champ de la science des données.

---

[72] DRUCKER, Johanna, 2010. 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. « The instrumental use towards specific ends and tasks that characterizes bureaucratic adoption of statistical methods and their graphic representation shifts the management of information from an intellectual to a political sphere. We can discern the ideological aspect of any scientific inquiry, but the applied use of information management makes use of the cultural authority in statistical graphics in a way that exceeds the qualified reservations of scientific method ». p. 15-16.

[73] TUFTE, Edward. 2001. *The Visual Display of Quantitative Information*, "Graphical Excellence." Cheshire, Connecticut: Graphics Press. p. 13.

[74] Les diagrammes en barre, par exemple, proviennent du champ des analyses et fonctions statistiques, tandis que les structures arborescentes sont le fait de la biologie évolutionnaire et de la généalogie.

[75] DRUCKER, 2010. « But the cultural history of information graphics, diagrams, maps, charts and other schematic images, is a rich field to mine for productive models within the horizon of the theoretical tools provided by the humanities ». p. 25.

Un changement épistémologique

Par le simple fait qu'elle permet d'embrasser du regard l'ensemble du phénomène qu'elle décrit, la visualisation de données peut-être considérée comme productrice de connaissances. Johanna Drucker décrit ainsi trois manières fondamentales par lesquelles l'image est amenée à devenir informative : « (…) 1) en offrant une analogie visuelle ou une ressemblance morphologique, 2) en fournissant une image visuelle d'un phénomène non-visible, ou 3) en fournissant des conventions visuelles pour structurer des opérations ou des procédures »[76].

Mais parallèlement à l'énonciation de ces différentes méthodes, ce qui caractérise la visualisation de données dans le cadres des humanités numériques, c'est un changement épistémologique fondamental. Là où Edward Tufte présupposait implicitement l'idée que « non seulement les données pré-existent à leur présentation graphique, mais aussi que les données ont une identité absolue en dehors de leur représentation »[77], Johanna Drucker considère au contraire que « l'epistémologie visuelle est basée sur une théorie plus radicale de la connaissance ».

> « Le concept radical de la subjectivité, et la nature co-dépendante de la connaissance et de l'interprétation, ont été essentielles à la physique quantique depuis près d'un siècle, mais aussi aux études cognitives depuis 50 ans. La "graphesis"[78] considère ces concepts comme fondamentaux et les utilise pour construire une théorie de la connaissance à travers l'attention portée aux formes graphiques et à ses nombreuses expressions ».

La visualisation des données n'existe donc pas pour représenter une information qui lui pré-existe, mais bien pour faire émerger une connaissance à partir de l'interprétation de l'observateur : pour le démontrer, Johanna Drucker se réfère en premier lieu à la physique quantique qui « suggère qu'un phénomène apparaît lorsqu'un observateur intervient dans un champs de potentialités »[79] . La visualisation incarne ainsi ce « champ de potentialités » dans lequel intervient un spectateur appelé à créer une connaissance par le biais d'un processus dynamique. L'auteur fait également références aux études cognitives, dont le psychologue gestaltiste Ernest von Glaserfeld peut être vu comme le précurseur, suggérant que « la cognition humaine émerge de manière dynamique dans une relation d'échange entre des capacités physiologiques et des stimulation circonstancielles dans un système continuellement changeant »[80].

Dès lors, la connaissance n'est plus transmise et révélée à un observateur passif, mais bien élaborée de manière dynamique par une interaction entre deux subjectivités, celle du créateur de la visualisation et celle de son spectateur : tout cela n'est guère éloigné, en un sens, de l'idée d'œuvre ouverte chère à Umberto Eco, selon laquelle « en réagissant à la constellation des stimuli, en essayant d'apercevoir et de comprendre leurs relations, chaque consommateur exerce une sensibilité personnelle, une culture

[76] *Ibid.* « Images embody information through three different models, each of which has a different structural relation to the referent. They can work 1) through offering a visual analogy or morphological resemblance, 2) through providing a visual image of non-visible phenomena, or 3) by providing visual conventions to structure operations or procedures ». p. 4.

[77] *Ibid.* « The assumption is not only that the data pre-exists the graphical presentation, but that the data have an absolute identity outside of their representation ». p. 24.

[78] Terme forgé en 1975 par Marie-Rose Logan, et désignant l'ensemble des réflexions produites à l'époque sur l'écriture et l'inscription.

[79] *Ibid.* « Rather than imagine discrete phenomena available for observation, or the subject-object relationship as a dialogue between two independent entities, the quantum theorist suggests that phenomena arise when an observer intervenes in a field of potentialities ». p. 28.

[80] *Ibid.* « Ernst von Glaserfeld's work in radical constructivism suggests that human cognition emerges dynamically in a relationship of exchange between physiological capabilities and circumstantial stimulation in a continually mutating system ». p. 27-28.

déterminée, des goûts, des tendances, des préjugés qui orientent sa jouissance dans une perspective qui lui est propre »[81].

Il reste à voir comment ces théories s'appliquent à la visualisation des données des bibliothèques.

L'exemple de l'Observatoire Bibliothèque

Le contexte de création de l'Observatoire

Le projet de l'Observatoire Bibliothèque (Library Observatory)[82] est né de l'effort commun de deux institutions : MetaLAB, le centre des humanités numériques de l'Université d'Harvard, d'une part, et la Digital Public Library of America (DPLA), d'autre part. Le MetaLAB se définit lui-même comme une « unité de recherche et d'enseignement (…) dédiée à l'exploration et l'expansion des frontières de la culture en réseau dans les arts et les humanités »[83]. Un des projets les plus intéressants du MetaLAB demeure sans doute l'Artefact de données ou Data Artifact, en raison notamment de sa réflexion approfondie sur la nature des données culturelles et leurs origines :

> « L'Internet inspire les bibliothèques, les archives, les musées et arboretums dans leur mouvement pour rendre leurs collections « ouvertes », « participatives » et « démocratiques ». Cet ensemble de valeurs, émergentes dans les cultures en réseau, se saisit d'institutions qui charrient depuis longtemps avec elles les legs de normes précédentes : la conservation, l'expertise, l'exhaustivité, l'excellence, et la commémoration. Dans certains cas, ces valeurs émergentes apportent des éléments aux plus anciennes ; dans d'autres cas, elles semblent en conflit. (…) Grâce à une attention critique portée au catalogage et aux schémas de classifications au travers de contextes institutionnels variés, l'Artefact de Données aura pour objectif d'historiciser les cultures de collecte et de comprendre les cultures matérielles et les valeurs intellectuelles qu'ils incarnent »[84].

Pour mettre au jour les valeurs, institutionnelles et politiques qui ont présidé à la création des données des collections et les éventuels conflits suscités par leur agrégation, il ne faudrait donc plus alors considérer les données comme « brutes », mais comme des médias, au même titre que les algorithmes qui permettent de les analyser.

Quel meilleur exemple pouvait-on choisir, dans ce contexte, que la DPLA, bibliothèque numérique américaine ? Dans un article intitulé « La chandelle de Jefferson »[85], Robert Darnton, éminent historien, directeur de la bibliothèque d'Harvard et co-fondateur de la DPLA, expliquait en effet les motivations qui

[81] ECO, Umberto. 1965. *L'œuvre ouverte*. Collection « Points », Éditions du Seuil, Paris. p. 17.

[82] Library Observatory, [sans date]. [en ligne]. [Consulté le 29 janvier 2014]. Disponible à l'adresse : http://www.libraryobservatory.org/

[83] About | metaLAB (at) Harvard, [sans date]. [en ligne]. [Consulté le 7 août 2014]. Disponible à l'adresse : http://metalab.harvard.edu/about/ « Metalab is a research and teaching unit at Harvard University dedicated to exploring and expanding the frontiers of networked cultures in the arts and humanities ».

[84] LOUKISSAS, Yanni, [sans date]. Data Artifacts Rising: Cultures of Collecting from Preservation to Participation | metaLAB (at) Harvard. [en ligne]. [Consulté le 19 mai 2014]. Disponible à l'adresse : http://metalab.harvard.edu/2012/12/data-artifacts-rising-cultures-of-collecting-from-preservation-to-participation/ « The Internet inspires libraries, archives, museums and arboreta to make their collection « open », « participatory », and « democratic ». This cluster of values emergent in networked cultures, is taking hold at institutions that carry long legacies of prior norms : preservation, expertise, comprehensiveness, excellence, and commemoration. In some cases, the emerging values adduce to older ones ; in other cases they seem to clash. (…) Through critical attention to cataloging and classification schemes across varied institutional contexts, Data Artifacts will historicize cultures of collecing and the understandings of material culture and intellectual value they embody ».

avaient présidé à sa fondation. Au gigantesque projet de numérisation porté par Google Book et ce qu'il était devenu, à savoir une opération de recherche transformée en « spéculation commerciale fondée sur la valeur de la base de données des livres »[86], Darnton oppose en effet le principe de bien commun de la connaissance, incarné par l'idéal Jeffersonien : « Qui reçoit une idée de moi, reçoit lui-même une instruction sans amoindrir la mienne ; de même que celui qui éclaire sa chandelle à la mienne reçoit de la lumière sans me plonger dans l'obscurité »[87].

Il s'agirait donc bien d'un conflit entre deux visions différentes de l'économie de l'information : d'une part celle, libérale, de Google, et d'autre par celle des « biens communs de la connaissance », portée en France par le collectif SavoirCom1, pour ne citer que cet exemple[88]. Ces deux conceptions donnent naissance à des données et des technologies de traitement différentes, à savoir Google Book et la DPLA, mais la dernière relève bien de cette inspiration vers l'ouverture, la participation et la démocratie décrite plus haut par MétaLAB, qui peut précisément amener des conflits avec d'autres valeurs ayant présidé à la création des données de la DPLA. En effet, contrairement à Google Book, explique Robert Darton, « la DPLA ne puisera pas dans une seule et gigantesque base de données. Il s'agira d'un système dit "distribué", qui agrégera les collections de multiples bibliothèques de recherche, musées et autres institutions ». Les diverses institutions qui ont fourni leur données de numérisation à la DPLA sont elles-mêmes en possession de ces héritages de normes et de valeur passées qui pourraient entrer en conflit avec les valeurs émergentes, propres à Internet, de la bibliothèque numérique américaine.

Or, ces conflits permettent de nous apporter des connaissances sur le contexte institutionnel des bibliothèques qui ont participé au projet porté par la DPLA. Or, tout l'enjeu est de pouvoir les mettre au jour.

Comment fonctionne l'Observatoire ?

> « En langage technique, un artefact de donnée est un objet produit par inadvertance au cours de processus humains d'organisation et de gestion. D'un point de vue culturel, un artefact est une fabrication située dans un contexte culturel. Enfin, d'un point de vue historique, un artefact est une trace évidente d'une rencontre à caractère médico-légale avec le passé. Jamais brute, toute donnée transporte les traces du travail humain, de ses interprétation et de ses valeurs »[89].

Matthew Battles et l'équipe de chercheurs à l'origine de l'Observatoire Bibliothèque ont la conviction que l'analyse et la visualisation de données rendent possible le repérage d'artefacts, ces « erreurs » de catalogage qui témoignent d'un conflit de classification entre deux institutions différentes à l'origine des données de la DPLA. Plus concrètement, l'Observatoire est une application conçue à partir de l'API (pour Application Programming Interface) fournie par la bibliothèque numérique américaine.

[85] Le débat. La chandelle de Jefferson, [sans date]. [en ligne]. [Consulté le 7 août 2014]. Disponible à l'adresse : http://le-debat.gallimard.fr/articles/2012-3-la-chandelle-de-jefferson/

[86] *Ibid*.

[87] *Ibid*.

[88] Nous renvoyons au mémoire de Clément Tisserand sur le sujet : TISSERANT, Clément, 2013. *Domaine public et biens communs de la connaisance*. Sous la direction de Cristina Ion. Disponible à l'adresse Web : http://www.enssib.fr/bibliotheque-numerique/documents/64245-domaine-public-et-biens-communs-de-la-connaissance.pdf

[89] BATTLES, Matthew. 2013. « Data artefacts : tracking knowledge-ordering conflicts through visualization. » dans INTERNATIONAL UDC SEMINAR, Slavić, Aida et UDC CONSORTIUM (THE HAGUE) (éd.), 2013. *Classification & visualization: interfaces to knowledge : proceedings of the International UDC Seminar 24-25 October 2013, The Hague, the Netherlands ; organized by UDC Consortium, The Hague*. Würzburg : Ergon. « This paper introduces the expression “data artefact” with the understanding that “artefact” has at least three meanings. In technical language, an artefact in data is an inadvertent product of human processes of organization and management. From a cultural perspective, an artefact is a designed object situated in a cultural context. Finally, from a historical perspective, an artefact is an evidentiary trace in a forensic encounter with the past. Never raw, all data carry traces of human labor, interpretations and values ». p. 244.

Elle permet de naviguer dans les collections à partir d'une visualisation représentant les collections sous forme d'une « carte arborescente » (en anglais « tree maps »). À un premier niveau, la carte arborescente permet de se faire une idée de la taille relative des contributions de chacune des institutions ayant participé à la bibliothèque numérique américaine, mais permet permet également de naviguer dans les collections de la DPLA (figure ci-dessous[90]).

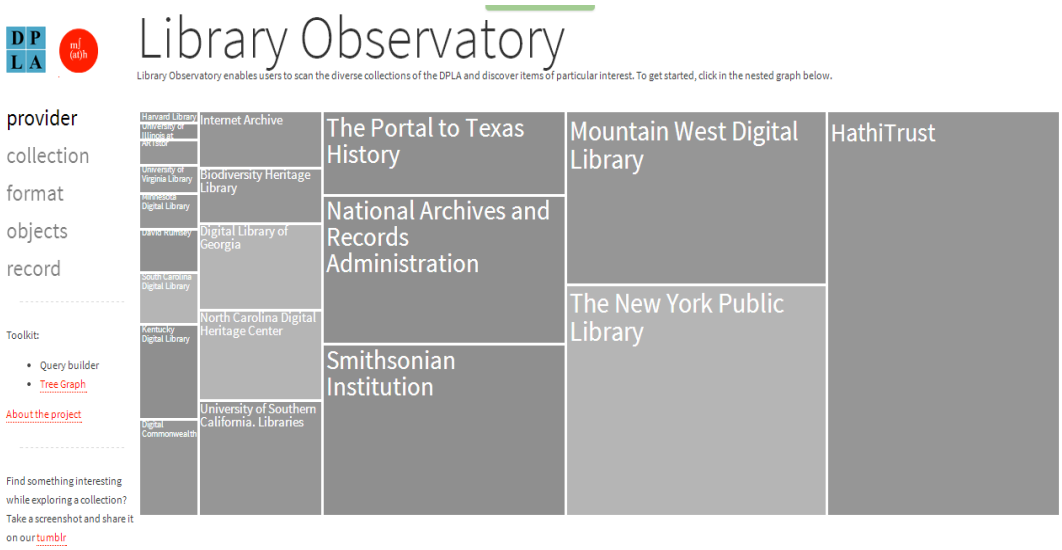


Figure 1 : Interface de l'Observatoire, montrant la taille relative des institutions ayant participé à la DPLA.

En effet, lorsque l'on clique sur l'un des carrés correspondant à un dépôt, une autre carte arborescente s'affiche et représente les différentes collections proposées par l'institution choisie puis, une fois choisie la collection, une troisième visualisation permet de choisir à l'intérieur de la collection un format souhaité, et ainsi de suite jusqu'au document à proprement parler[91] , chaque niveau indiquant une taille relative des objets visualisés. Ajoutons qu'une autre navigation, est proposée sous forme d'arbre, mais que le principe reste le même[92].

À partir de cette visualisation, l'Observatoire propose un tumblr, plate-forme de microblogage sur laquelle les utilisateurs de l'application sont invités à poster toutes les anomalies remarquées dans les cartes arborescentes ou les arbres simples, ainsi que les commentaires que leur inspire ces anomalies[93]. Un utilisateur a par exemple posté une capture d'écran montrant la carte arborescente des collections de la Smithsonian Institution au niveau de ses collections (figure ci-dessous). Cette visualisation montre une collection intitulée « type registre » (« type register ») qui, lorsqu'on cliquait dessus, ne menait à rien.

[90] mbattles_udcseminar2013.pdf, [sans date]. [en ligne]. [Consulté le 1 septembre 2014]. Disponible à l'adresse : http://www.udcds.com/seminar/2013/media/slides/mbattles_udcseminar2013.pdf

[91] *Cf* annexe, p. 100, figure 14.

[92] *Cf* annexe, p. 100, figure 15.

[93] Ce tumblr est actuellement disponible à l'adresse Web : http://libobserve.tumblr.com/.

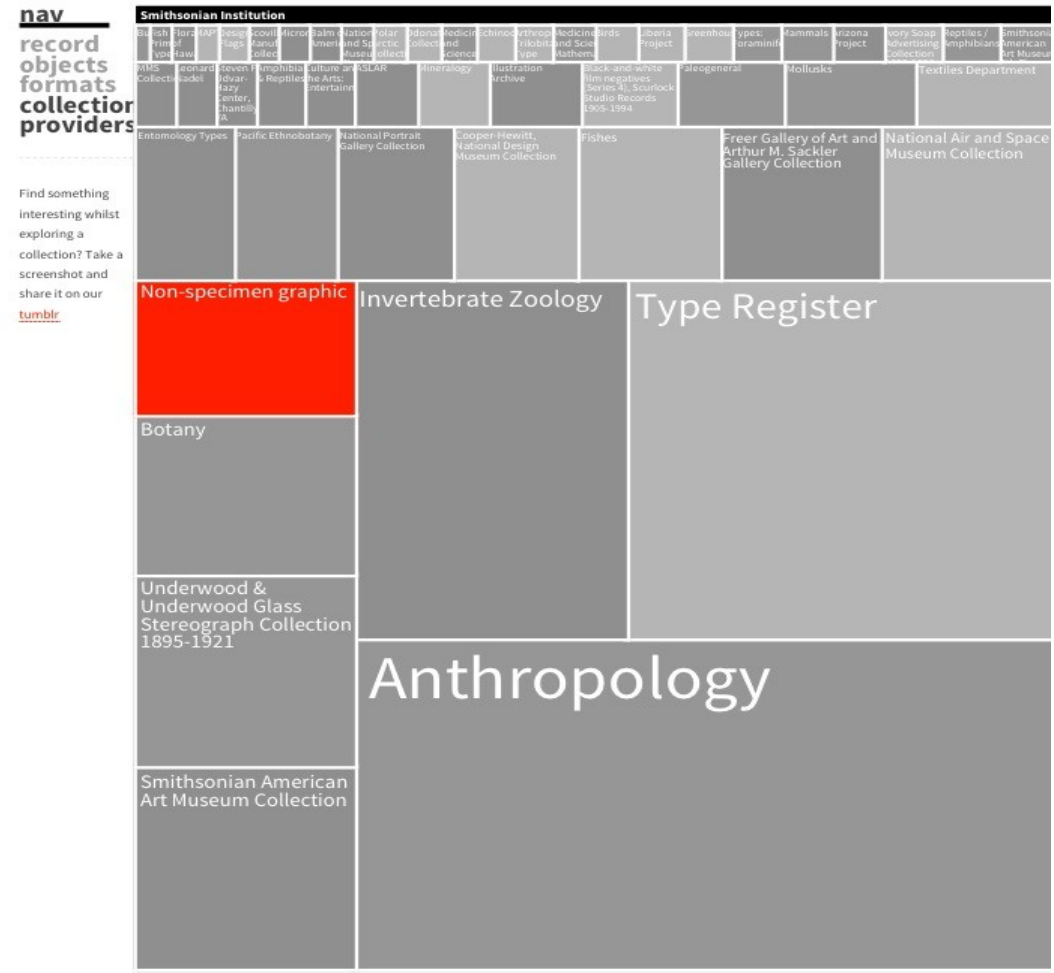


Figure 2 : Visualisation générée par un usager de l'Observatoire de la bibliothèque : anomalie du « type register » dans la classification arborescente de la Smithsonian Institution.

Après avoir fait remarquer que la carte arborescente révélait une tendance institutionnelle de la part de la Smithsonian Institution à privilégier l'anthropologie, Matthew Battles commente ainsi :

> « L'impasse où mène le type registre peut refléter une erreur dans l'ingestion de données ou des désaccords entre les modèles de données de la Smithsonian Institution et la DPLA. Les tailles variées des collections reflètent peut-être la signification relative de certains objets dans le programme de développement des collections – mais nous ferions bien de nous rappeler que la DPLA est une "méga-méta-collection" d'objets numériques, elle révèle non seulement les habitudes de collecte de ses institutions partenaires, mais aussi la concentration et la vitesse relative des efforts de numérisation de toutes ces institutions, tendances influencées non seulement par des prédilections de conservation ou d'université, mais aussi par les sources de financement, les préoccupation de préservation, et la susceptibilité à la numérisation présentée par de nombreux types d'objets »[94].

[94] BATTLES, 2013. « The dead-end at type register may reflect an error in the ingestion of data, or disagreemnets between th data models of the Smithsonian and the DPLA. The varied sizes of collections might reflect the relative significance of certain materials in the collections-development program – but we do well to remember that the DPLA is a « mega-meta collection » of digital objects ; it reflects not only the collecting habits of its partner institutions, but the focus and relative pace of digitization efforts across those institutions, patterns influenced not only by curatorial or scholarly predilections, but funding

L'Observatoire s'intéresse donc, au même titre que l'OCLC, à des questions relatives à la politique documentaire des établissements, et ce dans le cadre des bouleversements amenés par la numérisation massive. Si les méthodes sont radicalement différentes, elles sont à nos yeux complémentaires : les rapports de l'OCLC, d'une part, produisent des connaissances à une échelle globale. Ses observations sont de portées très générales mais permettent de se faire une première idée de ce que peut être une collection collective. Elle ne prend pas en compte, ou n'assume pas, la subjectivité de l'analyse et ne fait pas non plus appel à la subjectivité d'un observateur : les problèmes de normalisation des données qu'elle rencontre sont vus comme des obstacles devant être surmontés pour clarifier son propos. Au contraire, les observations faites au sein de l'Observatoire portent sur une échelle beaucoup plus restreinte et reposent sur la réaction subjective d'un utilisateur vis-à-vis de la visualisation des données de la DPLA : non seulement elle assume la subjectivité, mais elle en fait le point de départ de toute connaissance possible sur le contexte institutionnel des bibliothèques qui ont participé au projet. On voit donc là l'application des principes énoncés plus haut par Johanna Drucker : les conflits de classification impliqués par les formats d'origine très diverses des données sont perçus comme un outil indispensable de connaissance.

Conclusion : De la connaissance à la décision

Nous avons tenté, dans ce premier moment de notre réflexion, de faire le tour des différentes techniques employées à ce jour pour faire parler les données, depuis les statistiques inférentielles utilisées à l'occasion des enquêtes de publics, jusqu'à la visualisation interactives des métadonnées d'une gigantesque bibliothèque numérique telle que la DPLA, en passant par la confrontation de systèmes de recommandation et d'algorithmes de classement au sein de WorldCat et d'Amazon[95]. Mais, pourrait-on objecter, quel rapport peut-il y avoir entre la connaissance sur les usages des bibliothèques que pourraient prodiguer les enquêtes de publics et la connaissance institutionnelle que pourrait effectivement apporter la visualisation des métadonnées ? Peut-on véritablement mesurer les apports du Big Data et de la sciences des données en comparant des pratiques aux méthodes et aux objectifs fort différents ? Quel est, dans ce contexte, le réel apport des nouveaux outils par rapport aux pratiques déjà existantes permettant de connaître les bibliothèques ?

De fait, les bibliothèques n'ont pas attendu les nouvelles techniques apportées par les mégadonnées pour faire parler leur données. On trouve en effet dans le *Guide des études de publics en bibliothèque* un chapitre entier consacré à « la connaissance des publics via les données de la bibliothèque » qui développe « l'idée selon laquelle, avant même d'envisager la réalisation d'une enquête, les bibliothèques disposent elles-mêmes d'une multitude d'informations dont l'exploitation permet de fournir une connaissance riche et parfois unique des usagers et des usages dont elles font l'objet[96] ». Si donc les bibliothèques font déjà usage des données d'inscriptions pour « appréhender la capacité de l'établissement à susciter l'intérêt de la population qu'il dessert[97] », des données de portiques pour mesurer la fréquentation de la bibliothèque, du volume de non-inscription pour se

sources, preservation concerns, and the susceptibility of various kind of objects to digitzation ». p. 252.

[95] *Cf* annexe p. 98, figure 12, p. 99, figure 13.

[96] POISSENOT, Claude. « La connaissance des publics via les données internes de la bibliothèque » dans EVANS. 2011. p. 47.

[97] *Ibid*. p. 48.

faire une idée de « la capacité de l'équipement à fidéliser et donc à satisfaire ses usagers[98] », des données d'emprunts pour mesurer le taux de rotation de ses collections, etc., quels changements peuvent bien apporter le fait que ces données soient plus massives, que les techniques qui permettent de les appréhender soient plus performantes et que les résultats de ces analyses soient mobilisées dans le processus de prise de décision pour un établissement[99] ?

La réponse à cette question pourrait se trouver précisément dans les discours qui accompagnent aujourd'hui la science des données et le mouvement du Big Data dont elle est la traduction : jamais en effet le discours de l'innovation et de la « révolution » n'a davantage posé la question du statut ontologique des données. En clair, plus les Google et Amazon affirment avec force que les données générées par nos comportement passés permettent d'inférer sur nos comportement futurs, plus nous sommes amenés à nous poser la question de ce que sont les données et de ce que peut être un indicateur pour nous, chercheur ou décideur. En somme, plus nous réfléchissons sur les rapports entre les variables que nous produisons et ce qu'elles sont amenées à représenter à nos yeux, plus une prise de décision informée par ces variables requiert une recherche davantage approfondie sur la bibliothèque et son environnement.

[98] *Ibid*. p. 54.

[99] C'est déjà le cas, comme on pourra le constater dans la seconde partie de cette étude.

LES DONNÉES, UN ATOUT POUR LA GESTION D'UNE BIBLIOTHÈQUE ?

La production de connaissances sur un établissement est un atout pour son directeur, qui pourrait alors, par exemple, être en mesure d'améliorer son fonctionnement. Cette question de connaissances sur la bibliothèques entraîne donc naturellement une autre interrogation, qui pourrait être formulée de cette manière : dans quelle mesure les données peuvent-elles informer les décisions relatives à la gestion d'une bibliothèque ? Et si l'on inclut, dans ce que l'on entend par gestion d'une bibliothèque, la communication autour de son activité, quel peut-être à cet égard l'apport de la science des données, et notamment de la visualisation ?

Les réponses que nous tenterons de donner à ces questions pourront s'appliquer indifféremment, nous semble-t-il, aux bibliothèques universitaires et aux bibliothèques publiques : nous considérons en effet que la tâche d'évaluer un service, si elle sera dans un premier temps décrite en prenant le cadre universitaire, peut très bien se transposer à l'échelle d'une bibliothèque municipale. De même, le second chapitre de cette partie, qui se penchera sur un exemple de formation mis en place dans une bibliothèque de recherche aux États-Unis, peut s'envisager également dans le contexte publique : si la nécessité de mettre à disposition des usagers de la bibliothèque des personnes compétentes pour gérer les données issues de la recherche a motivé la création de la formation que nous allons décrire, il n'est pas évident que seul un public d'étudiants et de chercheurs, aujourd'hui, éprouve le besoin d'obtenir des données ainsi que les renseignements pouvant les accompagner. Enfin, si le dernier temps de ce chapitre porte spécifiquement sur la communication du bibliothécaire avec son élu, il nous semble que l'élu peut tout aussi bien être inter-changé dans notre propos avec le président d'université, dont le pouvoir de décision sur ses services de documentation a été renforcé par la loi sur l'autonomie des universités.

S'APPUYER SUR L'ANALYSE DE DONNÉES POUR ÉVALUER LA BIBLIOTHÈQUE...

Les bibliothèques françaises s'appuient sur une longue tradition d'évaluation. Il n'est qu'à prendre l'exemple de la création de l'Inspection générale des bibliothèques, en 1822, pour s'en convaincre : cette dernière était en effet chargée de mener des enquêtes ponctuelles sur le fonctionnement des bibliothèques afin de compléter les renseignements souvent lacunaires que les établissements de lecture devaient transmettre au ministère de l'Instruction Publique par le biais des rapports annuels adressés par les préfets et recteurs[100].

Dans ce contexte, il serait erroné d'affirmer que l'utilisation des données des bibliothèques françaises dans le but de piloter ces dernières serait un fait totalement nouveau : tout nous montre au contraire que cela a été pratiqué dès leurs origines, si l'on fait remonter ces origines aux confiscations révolutionnaires. Dès lors, l'analyse des données des bibliothèques, effectuée dans le contexte nouveau de la « science des données », n'aurait-elle rien de neuf à apporter ?

[100] ALONZO, Valérie, RENARD, Pierre-Yves (dir.). 2012. *Évaluer la bibliothèque*. Bibliothèques (Paris. 1978), 0184-0886. p. 38.

De la macro- à la micro-évaluation.

Mis en place depuis 2006, le système d'évaluation du SCD2 de Grenoble est remarquable par bien des aspects, notamment par la mobilisation de toute l'équipe du SCD autour de la mission – décrite dans une lettre de cadrage –, consistant à « doter la bibliothèque d'un outil d'aide à la décision centré sur la mesure de l'activité et des performances du service »[101], comme l'écrit Nadine Delcarmine :

> « Très rapidement, la réflexion sur le panel des indicateurs à suivre et sur la nécessité d'une large mobilisation des personnels dans des circuits de collecte et d'analyse efficaces a conduit le SICD à s'engager dans une démarche professionnelle hardie qui s'est traduite par la mise en place de plusieurs outils techniques produits en interne ou issus du monde de l'informatique décisionnelle (*business intelligence*) et, aussi souvent que possible, de la simplification et de l'automatisation des circuits »[102].

L'utilisation d'outils informatiques adaptés caractérise donc également le système d'évaluation mise en place à Grenoble. En premier lieu, il s'agit du référentiel des indicateurs, permettant de standardiser la collecte des données afin de permettre une utilisation à long terme de ces dernières ainsi que des comparaisons avec d'autres jeux de données. À cela s'ajoutent deux bases de données, une première destinée à la collecte des données qui ne sont pas issues du SIGB mais relevées manuellement par les agents du SCD et une seconde, « réplique du SIGB mise à jour continuellement » et qui fournit « les éléments statistiques sur les lecteurs, leur activité, le volume, la nature et les usages des collections imprimées »[103]. Pour couronner le tout, un outil de calcul, sous la forme d'une suite informatique appelée Cognos, doit permettre « d'élaborer des rapports et tableaux de bord préprogrammés par l'équipe du SICD »[104].

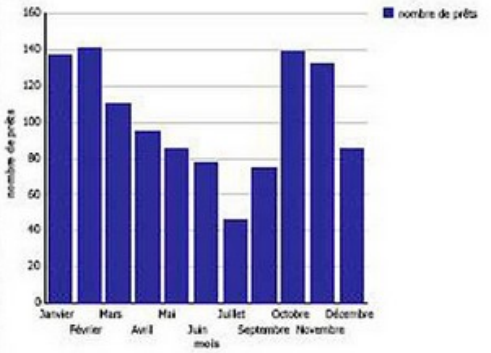
Il s'agit d'un outil technique en effet plutôt complexe, mais permettant cependant de simplifier la collecte et l'analyse mutualisées des données du SICD, les indicateurs ainsi construits devant permettre d'informer les décisions prises sur l'ensemble de l'établissement. Or, ces indicateurs, quels sont-ils ? Il s'agit de données pour l'essentiel quantitatives, comme on l'observe dans les différents tableaux de bords publiés en annexe du chapitre de Nadine Delcarmine[105] ainsi que dans la figure ci-dessous[106], permettant d'informer le bibliothécaire sur un certain nombre de points. La répartition des effectifs de la bibliothèque entre départements, par exemple, est un indicateur permettant de « mesurer l'impact du service public et de sa composante, la formation des lecteurs, sur l'activité globale de la bibliothèque »[107]. De même, « l'observation de l'évolution du nombre d'articles consultés par base, le calcul du coût d'une consultation, d'une recherche ou d'une session rapporté à une population d'utilisateurs donnée » permettent de « vérifier l'adaptation de l'offre documentaire ou du dispositif de formation à la recherche documentaire »[108].

[101] DELCARMINE, Nadine. « Tableaux de bord en bibliothèque » dans ALONZO et RENARD, 2012. p. 101.
[102] *Ibid*. p. 100-101.
[103] *Ibid*. p. 102.
[104] *Ibid*.
[105] *Ibid*. p. 104.
[106] DENNI, Gaëlle, 2010. Quatre catégories d'outils pour l'auto-évaluation au SICD2 de Grenoble. [en ligne]. 1 janvier 2010. [Consulté le 26 juillet 2014]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2010-04-0023-005
[107] DELCARMINE dans ALONZO et RENARD. 2012. p. 104.
[108] *Ibid*.

Prêts de la bibliothèque/du centre de documentation

EVOLUTION DANS L'ANNEE

succursale_du_pret	localisation_exemplaire	année civile	mois	nombre de prêts
8	8CER	2008	Janvier	133
		2008	Février	140
		2008	Mars	109
		2008	Avril	95
		2008	Mai	82
		2008	Juin	78
		2008	Juillet	46
		2008	Septembre	75
		2008	Octobre	139
		2008	Novembre	132
		2008	Décembre	85
8CER				1 114
8	8DPS	2008	Janvier	4
		2008	Février	1
		2008	Mars	1
		2008	Mai	3
8DPS				9
Total				1 123



REPARTITION PAR POLITIQUE DE PRETS

succursale_du_pret	localisation_exemplaire	Type de lecteur	nombre de prêts	pourcentage
8	8CER	8D	180	16,03%
		8LP	7	0,62%
		8M1	360	32,06%
		8M2	377	33,57%
		8P	136	12,11%
		8S	22	1,96%
		LP	32	2,85%
	8CER		1 114	
	8DPS	8D	1	0,09%
		8M1	3	0,27%
		8M2	5	0,45%

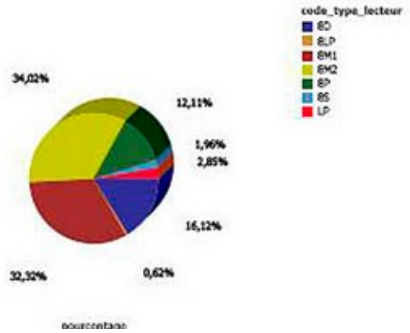


Figure 3 : Rapports statistiques de l'outil Cognos au SC2 de Grenoble

Mais ces données quantitatives constituant des indicateurs qui, comme le souligne Nadine Delcarmine, ne sont certes que des « points de repères qui doivent impérativement être resitués dans le contexte au moment de l'analyse »[109], suffisent-elles à informer sur l'état de la bibliothèque et sur son fonctionnement ? C'est la question que se pose Jamene Brooks-Kieffer, auteur d'un article sur la périlleuse artificialité des données des bibliothèques :

> « Les manières traditionnelles avec lesquelles les bibliothécaires rassemblent et traitent les données vont rarement jusqu'à l'emploi des techniques d'analyse, de traitement, ou d'exploration qui serait considérées comme une nécessité dans toute autre profession aussi riches en données que la nôtre. De telles techniques ne sont pas aisées à employer mais elles produisent des résultats remarquablement informatifs, si ce n'est parfois inconfortables. Mais à la place, nous préférons n'avoir affaire qu'à la signification superficielle et rassurante de nos données, nous reposant sur une prédominance des variables quantitatives et les conclusions simples et arithmétiques que nous pouvons en retirer. Ces conclusions sont rassurantes car elle ne mènent que rarement à des résultats inattendus »[110].

[109] *Ibid*. p. 103.

[110] BROOKS-KIEFFER, Jamene. « Yielding to persuasion : Library Data's Hazardous Surfaces » dans ORCUTT, Darby, 2010. *Library Data: Empowering Practice and Persuasion*. ABC-CLIO. « The traditional ways in which librarians gather and process data often stop short of the analysis, processing, or mining techniques that could be considered a necessity in any other profession as data-rich as ours. Such techniques are not easy to employ but they produce remarkably informative, if at times uncomfortable results. Instead, we prefer to deal with the surface, safe meaning of our data, relying on a predominance of quantitative variables and the simple, arithmetic conclusions we can draw from them. Those conclusions are safe because they seldom yield to unexpected results ». p. 3.

Jamene Brooks-Kieffer fait donc un constat assez dur, celui que les bibliothécaires et professionnels de la documentations, pourtant si désireux de piloter leur action par l'usage des données, ne savent cependant pas les analyser de manière approfondie et se contentent souvent de chiffres et de simples opérations arithmétiques (pourcentages, minimum, maximum, moyennes, etc...) pour informer leur décision. Mais comment pourrait-on s'attendre à ce qu'il en soit autrement, continue-t-elle, alors que les agences nationales américaines elles-mêmes désignent sous le terme de « statistiques » une « collection de données quantitatives portant sur un sujet particulier » :

> « Lorsque les agences nationales elles-mêmes envoient ce message aux bibliothèques américaines que les seules statistiques significatives sont des collections de données numériques et de simple calculs effectués sur ces données, devrait-on s'attendre à ce que les bibliothèques elles-mêmes pensent ou agissent de manière différente »[111] ?

Il nous semble d'ailleurs que ce constat portant sur l'environnement institutionnel des bibliothèques américaines s'applique aussi bien à celui des bibliothèques françaises, l'observatoire de la lecture publique étant décrit par Valérie Alonzo comme un « réservoir statistiques (…) d'une grande richesse (dans la limite de la complétude et de l'exactitude des réponses apportées aux enquêtes) » permettant « d'exploiter les statistiques de façon synthétiques (calcul de ratios, de valeurs moyenne ou médiane) (…) »[112].

Cependant, quelle méthode d'analyse Jamene Brooks-Kieffer propose-t-elle à la place des traditionnelles données quantitatives et des indicateurs qui tiennent lieu pour nous de « statistiques » ? Pour répondre à cette question, l'auteur commence par opposer la macro-évaluation, dont l'objet est de s'intéresser à ce qu'un ensemble de variables peut dire d'un organisme, à la micro-évaluation qui se concentre elle sur « la manière dont ce jeu de variables est affecté par d'autres ensemble de variables »[113]. Pour illustrer ce propos, Brooks-Kieffer prend l'exemple d'une situation particulière, celle où un directeur d'une bibliothèque universitaire cherche à savoir dans quelle mesure les services de prêt répondent aux besoins en documentation des usagers distants de la bibliothèque. Pour répondre à cette question, une macro-évaluation se pencherait sur un indicateur possible de performance, à savoir le nombre total de documents empruntés, distribué en fonction des codes postaux des usagers. Nous ne sommes donc guère loin des indicateurs construit dans le cadre de la LOLF, où par exemple l'addition du nombre de prêts, du nombre de documents téléchargés, du nombre de documents communiqués sur place et du nombre de prêts PEB doivent renseigner sur l'usage des collections et imprimées et numériques de la bibliothèque.

Mais l'analyse ne doit pas s'arrêter là, écrit Brooks-Kieffer :

> « Pour déterminer, par exemple, pourquoi les usagers distants empruntent des documents à un taux de 75% inférieur à celui des usagers locaux, nous avons besoin de conduire une micro-évaluation. Nous pouvons rassembler d'autres données quantitatives et qualitatives provenant du SIGB et des usagers distants eux-mêmes afin d'étendre notre analyse de départ. La pratique de la micro-évaluation exige que nous examinions aussi d'autres enjeux qui ont un ef-

[111] *Ibid*. « When national agencies send the message to U.S. Libraries that meaningful statistics are collections of numeric data and simple calculations performed on that data, should we expect libraries themselves to think or act any differently » ? p. 7.

[112] ALONZO, RENARD. 2012. p. 107-108.

[113] ORCUTT, 2010. « Where macroevaluation is concerned with what a set of variables says about an organization, microevaluation addresses how that set of variables is impacted by other sets of variables and why the organization behaves as it does under the influence of these variables ». p. 10.

> fet sur les interactions des usagers distants avec la bibliothèque : les politiques mises en place au sein de l'établissement, les contraintes horaires, les contenus des cours, etc. S'il est vrai que cela complique beaucoup l'analyse, la micro-évaluation fournit une réponse plus complète à nos questions en tentant de prendre en considération des facteurs situés en dehors de la portée initiale des données »[114].

Mesurer par le biais d'autres jeux de données l'impact de certains facteurs sur les indicateurs observés, c'est donc le but de la micro-évaluation. Nous pourrions nous pencher maintenant sur quelques exemples qui illustrent peut-être davantage les idées développées par Jamene Brooks-Kieffer.

Quelques exemples innovants d'analyse des données en bibliothèque.

Un premier exemple qui, à nos yeux, illustre une plus grande souplesse et une plus grande profondeur d'évaluation que nos traditionnels tableaux de pilotage, nous semble être une étude conduite en 2013 à la bibliothèque de la faculté du New Jersey[115]. Cette étude a attiré notre attention pour trois raisons qui sont intrinsèquement liées à ses qualités méthodologiques.

En premier lieu, il ne s'agissait pas d'une étude visant simplement à produire des connaissances sur la bibliothèque, mais bien à évaluer l'activité d'un établissement dans le but d'influer directement sa gestion : en l'occurrence, il s'agissait de mesurer la pertinence des acquisitions les plus récentes de la bibliothèque pour ensuite influer sur la politique documentaire de l'établissement[116]. Deuxièmement, les auteurs ont choisi pour répondre à leur objectif d'interroger la corrélation de trois variables différentes : les acquisitions récentes, les circulations récentes, et les demandes de prêt entre bibliothèques également récentes. À cela s'ajoute la volonté de choisir une échelle restreinte en divisant les variables par groupes de lecteurs[117].

Enfin, l'intérêt principal de ce travail réside sans doute à nos yeux dans le fait que les auteurs on voulu réfléchir, du début jusqu'à la fin de leur étude, sur les rapports qu'entretenaient les variables choisies pour remplir leur objectif d'évaluation avec la réalité qu'elles étaient censées représenter :

> « Ce faisant, l'étude nous a conduit à envisager de quelle manière les données pouvaient nous aider à définir et à répondre aux questions fondamentales du développement des collections : sélectionnons-nous ce dont les usagers ont besoin ? Les données d'usages peuvent-elles nous aider à cerner ces besoins ? Avons-nous suffisamment rendu service à la fois à notre faculté et à nos étudiants ? Les demandent de PEB représentent-elles des failles dans les collections ou bien des désirs des usagers qui iraient au-delà de la portée de nos politiques documentaires actuelles ?

[114] *Ibid*. « To determine, for instance, why patrons check out items at a rate 75 percent less than that of local patrons, we need to conduct a micro-évaluation. We can gather other quantitative and qualitative data from the ILS and from distance patrons themselve to expand our original analysis. The pratcice of micro-evaluation requires that we also examine other issues that affect distant patron's interactions with the library : library policies, scheduling constraints, course content, and so forth. While complicatin the analysis a great deal, micro-evaluation provides amore complete answer to our question by attempting to consider factors outside the data's scope ». p.11.

[115] E.LINK, Forrest, TOSAKA, Yuji, WENG, Cathy. « Mining and Analyzing Circulation and ILL Data for Informed Collection Devlopment. » Preprint à paraître dans *College & Research Libraries*, 2015. Microsoft Word - Link-Tosaka-Weng.docx - crl14-632.full.pdf, [sans date]. [en ligne]. [Consulté le 8 décembre 2014]. Disponible à l'adresse : http://crl.acrl.org/content/early/2014/10/20/crl14-632.full.pdf

[116] *Ibid*. « By conducting similar evidence-based evaluation of library use patterns, we believe that academic libraries should be able to create effective feedback mechanisms to monitor and inform their collection development practices to better meet the changing needs of their user populations ».

[117] *Ibid*. « (…) relationships among recent acquisitions, circulation, and ILL borrowings data need to be examined more carefully to determine subject strengths and weaknesses in relation to the total user demand for library materials. Moreover , because academic libraries serve different user populations (e.g., undergraduate, graduate students, faculty), it is also essential that effort be made to disaggregate allt hese data sets and analyse them on a smaller scale to examine the effectiveness of collection development activities for different user categories ».

En contribuant à fournir des aperçus nouveaux sur ces questions, l'objectif final de cette étude était de renouveler le dialogue entre les acquéreurs et les usagers et redonner de l'énergie à la conception du développement des collections de la bibliothèques »[118].

D'une certaine manière, les auteurs de l'étude semblent admettre que les variables qu'ils ont choisies, à savoir le trio usage = satisfaction des besoins, PEB = insatisfaction des besoins, satisfaction des besoins = efficacité des politiques d'acquisition, ont un caractère entièrement construit et échouent sans doute pour une grande part à rendre compte du bon fonctionnement ou non de la politique documentaire de l'établissement. Mais ils reconnaissent en même temps que le grand mérite de cette réflexion menée en continu est d'avoir su rétablir un dialogue entre les acquéreurs et leur public, ce qui, d'une certaine manière, constitue peut-être un résultat plus intéressant que l'évaluation des collections en elle-même.

L'exemple de la bibliothèque de la faculté du New Jersey est cependant un exemple d'évaluation ponctuelle des activités de l'établissement. Si l'on voulait examiner un dispositif d'évaluation systématique, l'exemple le plus accompli est sans doute celui de la « ferme de données » de la bibliothèque universitaire de Pennsylvanie[119]. Ce projet de « ferme » est en effet né de la frustration éprouvée par les bibliothécaires chaque fois qu'ils tentaient de déceler de manière significative les comportements des usagers de leur bibliothèque à partir du flot de données d'usage des ressources électroniques dont ils disposaient. Ne contenant initialement que des données de log, la ferme a finit par s'étendre à toute sorte d'autres données provenant de sources multiples, l'aspect le plus intéressant de ce gigantesque entrepôt résidant sans doute dans un outil sur mesure de publication de rapports permettant d'aller au-delà des simples rapports sommaires que l'on peut observer, par exemple sur l'outil Cognos.

Prenons ainsi l'exemple d'une enquête annuelle menée dans la bibliothèque universitaire, intitulée « Qui pose des questions et où »[120] ? :

> « Durant quatre cycles par an, d'une durée d'une semaine chacun, la bibliothèque collecte des données sur diverses sortes de questions posées à des points de services ou des bureaux de bibliothécaire. En même temps que les questions, nous comptons également le nombre de documents empruntés et rangés, de même que le nombre de sorties de nos différents sites sur le campus. Nous avons remarqué que ces données sont corrélées de manière très significative, et le renouvellement annuel des comptes permet d'observer régulièrement les relations qui existent entre elles. Ces données sont utilisées pour donner des estimations, comme par exemple la fréquence des usages sur place des collections. Mieux, les comptes nous permettent d'évaluer la distribution, la variation et les changements dans les services de référence dispersées dans les communautés que nous desservons et leur bibliothèque »[121].

[118] *Ibid*. « In so doing, the study has led us to consider how usage data could help define and answer the fundamental questions of collection development : are we collecting what our user need ? Can usage data help us pinpoint these needs ? Have we sufficiently served both our faculty and students ? Do ILL requests represent collections failures or user wants beyond the scope of current collection policies ? By helping to provide fresh insights into these questions, the ultimate goal of this study was to refresh the dialogue between selectors and users and to re-energize library colection development thinking ».

[119] Penn Library Data Farm, [sans date]. [en ligne]. [Consulté le 13 mai 2014]. Disponible à l'adresse : http://datafarm.library.upenn.edu/

[120] Penn Library - Graduate Student Workshops, [sans date]. [en ligne]. [Consulté le 16 août 2014]. Disponible à l'adresse : http://datafarm.library.upenn.edu/desksurvey/index.html

[121] *Ibid*. « In four, week-long cycles each year, the Library collects data on various kinds of questions asked at service points and librarian offices. Along with questions, we count the number of items circulated and reshelved, as well as the number of exits from our campus locations. We have found that these data are highly correlated, and the annual renewal of counts helps to monitor the relationships between them. These data are used to estimate outputs, such as the

Voilà donc, à ce qu'il nous semble, un exemple approfondi d'évaluation : par un dispositif de collecte systématique de jeux de données hétérogènes, la ferme des données permet de corréler plus facilement des variables de natures assez différentes, et peut-être à un niveau de précision plus fin que ce qui est actuellement permis par les outils utilisés aujourd'hui en bibliothèque.

Penser les données des bibliothèques non comme des indicateurs mais comme des symboles de son activité

Dans son article sur la « dangereuse superficialité des données des bibliothèques »[122], Jamene Brooks-Kieffer tire une conclusion qu'il nous a paru intéressant de souligner :

> « (…) L'expression à la mode de « prise de décision fondée sur les données » fait faussement croire à de nombreux bibliothécaire qu'ils agissent correctement en fondant leur décisions sur la quantité d'une chose ou le pourcentage de telle autre chose. (…) La prise de décision fondée sur les données est une approche simpliste d'un problème complexe. Elle contourne toutes les formes les plus basiques d'analyse des données et ignore la possibilité que les décisions devraient prendre en considération des facteurs autres que des données quantitatives »[123].

Il nous semble que ce que l'auteur critique principalement n'est pas tant l'utilisation de variables quantitatives pour informer les décisions concernant la bibliothèque que la trop grande confiance accordée à ces variables, du fait de la volonté de certitude qui sous-tend le pilotage d'un établissement. Or, peut-être serait-il nécessaire que toute personne amenée à travailler avec les données des bibliothèques prenne conscience de ce que sont par nature les données : non des reflets du réel mais davantage des fragments de ce dernier dont le sens, jamais fixe ni certain, est sans cesse à construire.

Peut-être serait-il opportun, pour saisir ce caractère instable et faussement miroitant des données, de rapprocher ces dernières du statut des archives aux yeux de l'historien, notamment tel que l'expose Arlette Farge. Celle-ci décrit en effet la surveillance policière du Paris du XVIII[e] siècle : par un dispositif pyramidal de « mouches », le pouvoir monarchique a cherché à se tenir informé au plus près de l'opinion parisienne, notamment dans le but de prévenir d'éventuels soulèvements... Mais à lire ces paroles captées, comme la NSA capte aujourd'hui les données des télécommunications, on ne peut s'empêcher de sentir toute la vanité de cet effort de surveillance. Jamais en effet la rue n'a pu être entièrement saisie par le pouvoir royal, pour la simple raison que le système de surveillance masquait, par son dispositif même, la réalité des comportements quotidiens des administrés : « les formes mêmes de l'organisation policière sont construites autour de cette nécessité quotidienne de tout savoir et tout entendre, et le classement des archives du lieutenant général traduit cette préoccupation forcenée pour le détail et le goût de chaparder sans vergogne les paroles prononcées au hasard des conversation publiques »[124].

---

frequency of in-house collection use. More important, the counts allow us to assess the distribution, variation, and change in reference services across the communities we serve and their libraries ».

[122] ORCUTT. 2010.

[123] *Ibid*. « First, the faddish phrase "data-driven decision-making" misleads many librarians into believing that they are acting correctly by basing their decisions on how many of something or what percentage it is of some other thing. Yes, these are decisions based on data analysis, but I hope that I have shown that relying solely on such a minimal form of data analysis ofr organizational assessment is like trying to decode the human genome with a pocket calculator. Data-driven decision-making is a simplistic approach to a complex problem. It often bypasses all but the most basic forms of data analysis and ignores the possibility that decisions should consider factors other than quantitative data. In Lancaster's usage, this is decisions-making based solely on macroevaluation ». p. 12.

[124] FARGE, Arlette., 1997. *Le Goût de l'archive*. [Paris] : Seuil. p. 126.

Ainsi ne nous a-t-il pas paru primordial de traiter la question de l'enjeu d'atteinte à la vie privée que comporterait le Big Data. Les données, au même titre que les archives, ne permettent pas réellement de saisir la réalité des individus, et encore moins la réalité d'une bibliothèque :

> « L'archive pétrifie ces moments au hasard et dans le désordre ; chaque fois, celui qui la lit, la touche ou la découvre est d'abord provoqué par un effet de certitude. La parole dite, l'objet trouvé, la trace laissée deviennent figures du réel. Comme si la preuve de ce que fut le passé était enfin là, définitive et proche. Comme si, en dépliant l'archive, on avait obtenu le privilège de 'toucher le réel'. Dès lors, pourquoi discourir, fournir de nouveaux mots pour expliquer ce qui tout simplement gît déjà sur les feuilles ou entre elles »[125].

Dans une certaine mesure, les données se comportent de la même manière que les archives, donnant une première impression de réel qui tendrait à disqualifier le travail de l'historien : si les données parlent d'elles-mêmes, à quoi bon les faire parler ? De fait, s'il est une chose que la lecture attentive des archives, notamment judiciaires, nous apprend, c'est que le discours capté des justiciables est un discours construit pour la circonstance, et fonction des stratégies individuelles de chacune des parties, qu'il s'agisse de plaignants, de défendeurs ou des juges. Or, il ne nous semble pas que les données soient d'une nature fondamentalement différente puisque également dépendantes des circonstances et du contexte qui les ont vu naître : les données de circulation, d'acquisition, de fréquentation, etc., ne sont-elles pas construites elles aussi dans le contexte bien particulier de l'activité d'un établissement et non dans le but de rendre compte de cette dernière ? De ce fait, l'idée de faire parler les données ne serait pas autre chose que d'essayer de penser l'invention d'un langage qui s'adapte à elles, et qui, si l'on voulait reprendre les mots d'Alain Corbin, « autorise une quête en profondeur sans que le chercheur prétende [en] épuiser le sens[126] ». Pour l'heure, la visualisation est sans doute le langage qui le mieux, cherche à appréhender le caractère social et construit des données. Mais il n'est pas dit qu'il soit impossible à l'avenir d'intégrer cet aspect dans d'autres langages, tels que les algorithmes et l'apprentissage automatique.

Par ailleurs, cette réflexion autour du langage des données impliquerait nécessairement cette autre idée sous-jacente selon laquelle faire parler les données pose la question du réel et de ses représentations. Par là, il devient pertinent d'interroger la notion d'indicateur, notamment dans le contexte de la gestion d'une bibliothèque. En effet, si, comme l'affirme wikipédia, « l'utilité d'un indicateur dépend d'abord de sa capacité à refléter la réalité »[127], peut-être serait-il plus honnête de lui préférer la notion de symbole, telle que décrite par Paul Tillich[128] : à la fois représentations conventionnelles d'une réalité et éléments participatifs de la réalité qu'elles désignent, le travail sur les données permet d'ouvrir à des niveau de connaissance et de réalité qui autrement resteraient inaccessibles. Plus concrètement, travailler sur des variables qui seraient désignées comme symboliques plutôt qu'indicatives de la bibliothèque et de son activité permet de ne pas verrouiller leur signification qui resterait ainsi à construire de manière collective par les professionnels de l'établissement qui les a produit.

---

[125] *Ibid*. p. 18.

[126] ALAIN, Corbin, 1991. Arlette Farge, « Le goût de l'archive ». *Annales. Économies, Sociétés, Civilisations*. 1991. Vol. 46, n° 3, p. 595-597.

[127] Indicateur, 2014. *Wikipédia* [en ligne]. [Consulté le 9 novembre 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Indicateur&oldid=106207898. Page Version ID: 106207898

[128] TILLICH, Paul et GOUNELLE, André, 2012. *Dynamique de la foi*. Genève; Québec; [Paris] : Éd. Labor et fides ; les Presses de l'Université Laval ; [diff. les Éd. du Cerf]. p. 47.

Par leur nature ontologique, les métadonnées participent pleinement de ce caractère symbolique : l'essai intitulé *The life and death of data*[129] cherche précisément à mettre en lumière l'instabilité essentielle des métadonnées. Ainsi, de même que les symboles vivent et meurent avec les communautés qui les ont créés, de même les métadonnées vivent et meurent avec les contextes sociaux et culturels qui les ont vu naître. De ce fait, il nous semble que c'est bien reconnaître le caractère symbolique des données que de proposer de privilégier la visualisation comme moyen d'accéder aux connaissances : l'apprentissage d'un tel langage reste aujourd'hui à inventer et à mettre en place pour les professionnels des bibliothèques.

DST4L : UN EXEMPLE DE FORMATION SPÉCIALEMENT CONÇUE POUR DES BIBLIOTHÉCAIRES.

La DST4L (Data Scientist Training for Librarian) est une formation à l'analyse des données spécifiquement destinée aux bibliothécaires. Elle a été mise en place au sein du Harvard-Smithsonian Center for Astrophysics par Christopher Erdmann, directeur de la John G. Wolbach library. Nous nous proposons ici d'en décrire les caractéristiques principales avant d'étudier les projets mis en place en son sein, notamment ceux impliquant la visualisation des données.

Contexte et objectifs de la formation

Dans la présentation qu'il fait de sa formation[130], Christopher Erdmann insiste en premier lieu sur l'environnement institutionnel dans lequel évolue sa bibliothèque à savoir, d'une part, la communauté d'astronomes que celle-ci dessert et, d'autre part, la base bibliographique de la NASA, Astrophysics Data System (ADS). Cet environnement institutionnel met en évidence la spécificité des bibliothécaires de la Wolbach library, dont le rôle est de faciliter la réutilisation des données produites par l'activité de recherche astronomique :

> « Beaucoup de ce que nous, bibliothécaires, faisons, permet de faciliter la recherche et la découverte d'objets dans l'ADS, mais ce qui est plus important, c'est que nous générons un grand nombre de liens vers les données que les astronomes utilisent de manière quotidienne. Cette activité de gestion permet également de mesurer les performances des télescopes et instruments. Je pense que ce genre d'activité est une condition fondamentale pour centrer une bibliothèque sur les données »[131].

D'une certaine manière, nous pourrions dire que les bibliothèques d'astronomies disposent d'une relative avance dans la mise en place de service de gestion et d'ouverture des données de la recherche, comme en témoigne l'exemple français du centre des données de Strasbourg, cité par Rémi Gaillard dans son mémoire sur l'Open research data[132]. Or, la mise en place de ces services au sein des bibliothèques de recherche nécessite une formation adéquate des bibliothécaires, afin de pouvoir se « plonger dans les données, manipuler le cycle de vie de la donnée de recherche et se frotter au milieu

[129] The Life and Death of Data, [sans date]. *op. cit*.

[130] ERDMANN, Christopher, 2014. Teaching librarians to be data scientists. *Information outlook* [en ligne]. mai-juin 2014. Vol. 18, n° 3. [Consulté le 17 août 2014]. DOI 10.5281/zenodo.11217. Disponible à l'adresse : https://zenodo.org/record/11217/files/DataScientistTraining.pdf

[131] *Ibid*. « Much of what we librarians do helps facilitate search and discovery in the ADS, but more importantly, we generate many of the data links that astronomers use on a daily basis. This curation activity also supports analyses of how telescopes and instrumentation are performing. I believe this type of work forms the back-drop of the data-centric library ». p. 21.

[132] GAILLARD. 2013. p. 71.

de la "science des données" »[133]. Pour pouvoir développer l'Open Research Data dans les bibliothèques universitaires, il faut donc d'abord envisager des programmes de formation innovants, tels qu'Immersive Informatics, « un programme pilote anglo-australien élaboré par les université de Melbourne et de Bath » dont le but est d'apprendre à des bibliothécaires à gérer un jeu des données « en vue de sa conservation et de sa diffusion future »[134].

En évoquant ce contexte d'ouverture des données de la recherche, nous souhaitons souligner que la formation des bibliothécaires à l'analyse des données, dont nous avons démontré plus haut l'intérêt pour l'évaluation et la gestion des bibliothèques, s'inscrit dans un mouvement plus global, allant des sciences sociales aux sciences dures en passant par les Humanités Numériques. L'une des conséquences d'un tel mouvement pourrait être d'ouvrir la profession sur la gestion des données et les connaissances scientifiques qui l'accompagne. La formation des professionnels à la gestion et à l'analyse des données pourrait donc bénéficier du mouvement d'ouverture des données de la recherche. Bien sûr, on pourrait objecter à cela que ce mouvement ne concernerait que les professionnels appelés à travailler dans des bibliothèques de recherche. Mais en réalité, nous pourrions dire avec Lynda Kellam et Katharyn Peter que ce mouvement touche tous les types de publics et, avec eux, tous les types de bibliothèque. Il y a en effet deux facteurs a prendre en considération aujourd'hui : d'une part le fait qu'Internet a rendu l'accès et la circulation de jeux de données plus aisés et, d'autre part, l'essor de l'utilisation d'outils abordables tels que les tableurs Excel. Ces deux facteurs participent de fait à créer une culture du nombre de plus en plus accessible et partagée que les bibliothèques, mêmes publiques, pourraient être amenées à prendre en compte :

> « Avec ces changements dans l'accès aux données numériques, les bibliothécaires ont pris une place centrale dans l'aide aux usagers. Notre activité principale a peut-être été le mot écrit mais la montée des formats et des fichiers numériques a fait émerger un rôle nouveau pour la bibliothèque, un rôle qui soutient l'information sous toutes ses formes – depuis le mot écrit, jusqu'à l'image numérique, en passant par l'échantillon circulant en streaming et le fichier de données quantitatives. (…) Il se peut que les usagers n'associent pas immédiatement les ressources quantitatives à la bibliothèque mais de plus en plus de bibliothèques et de bibliothécaires sont appelés à acheter, communiquer et archiver ces ressources »[135].

La donnée est donc susceptible d'intéresser tout type de public et c'est bien à ce titre que Kellam et Peter en appellent à la formation de « donnéethécaires » (data librarians) dont le rôle serait de sélectionner, rendre disponible et promouvoir des jeux de données[136]. Dès lors, la difficulté réside dans le fait qu'à quelques expressions près, les bibliothécaires sont traditionnellement peu formés dans le domaine des statistiques et de la gestion informatisée des données.

L'objectif de la DST4L, néanmoins, consiste essentiellement à permettre aux bibliothécaires de nettoyer et rendre visible les jeux de données afin d'en faciliter la découverte, et non spécifiquement de former à l'analyse des données : en réalité,

[133] ERDMANN. 2014. p. 21.

[134] GAILLARD, 2013. p. 72.

[135] KELLAM, PETER, 2011. « With these changes in the access to numeric data, librarians have become central participants in assisting users. Our traditional focus may have been on the written word, but the rise in digital formats and files has carved out a new rôle for the library, one that supports information in all its forms – from the written word, to the digital image, the streaming media sample, and the numeric data file. Moreover, our promotion of information literacy and emphasis on information-literate users means we need to pay attention to all types of information sources, even the non-textual. Users may not immediately associate numeric data sources with the library, but inreasingly libraries and librarians are being called upon to purchase, support and archive these sources ». p. 2.

[136] Pour la définition d'un jeu de données, *Cf* introduction de cette étude, p. 13, note n°7.

il s'agit essentiellement d'alléger le travail des « data scientists » en effectuant toutes les tâches qui se situent en amont de l'analyse et qui occupent cependant 80% de leur temps. Pour autant, l'analyse et la visualisation des données constituent une grande partie du programme de la formation[137], puisqu'on peut y observer que pas moins de sept séances sur quinze sont consacrées aux statistiques, à la programmation et à la visualisation des données. Et nous aimerions souligner qu'en définitive, Christopher Erdmann insiste moins sur l'apport des bibliothécaires ainsi formés à la communauté qu'ils desservent que sur ce que cette formation leur a permis de faire afin d'améliorer les services de leur établissement :

> « Un autre objectif de la DST4L était de renouveler les compétences des bibliothécaires, et beaucoup de participants utilisent désormais leurs nouvelles capacités. Par exemple, Veronica Downey a automatisé certaines tâches de la bibliothèque en utilisant Python, Alex Holachek aide l'ADS à améliorer ses outils de visualisation et Katie Frey cherche à introduire les technologies sémantiques en astronomie »[138].

Mais s'il est vrai que ce type de formation s'avère nécessaire, est-il réaliste et faisable de former des bibliothécaires à la programmation et de leur faire acquérir une expertise en analyse des données ?

« Comment dompter les données bibliographiques »[139] ?

Au sein de la formation mise en place par Christopher Erdmann, certains participants ont été associés à un projet dont l'objectif était d'améliorer le fonctionnement de certaines tâches de l'ADS. Il s'agissait en effet de repérer dans l'immense bibliothèque numérique de l'Internet Archive (IA) les documents déjà possédés par l'ADS et ceux qui ne l'étaient pas, ce afin de compléter les collections numérique de l'ADS en pointant vers les ressources de l'Internet Archive là où celles-ci faisaient défaut[140]. Le projet a d'abord subi quelques modifications quant à ses objectifs, du fait de l'incompatibilité des formats de données entre l'ADS, qui décrit les documents à l'échelle d'un article, et l'IA, qui les décrit à l'échelle d'un titre de revue. Les responsables du projet ont donc décidé de se contenter dans un premier temps de retrouver des correspondances entre les données bibliographiques de l'IA et celle de l'ADS seulement pour les monographies, où les formats de données étaient à peu près similaires.

Les membres du projet ont ensuite commencé par extraire de l'ADS les données correspondant aux monographies d'astronomie dont les notices se trouvent dans la base bibliographique de la NASA, en se contentant des données de titre, auteur et date de publication. De là, ces données n'étant pas « propres » du fait du caractère composite de l'ADS qui regroupe plusieurs institutions différentes, un outil gratuit en ligne, OpenRefine, a été utilisé pour repérer et corriger automatiquement les erreurs de frappe ou de catalogage dans les données fraîchement récupérée. OpenRefine a ensuite été de nouveau utilisé pour construire des requêtes sur mesure afin de fouiller les données de l'Internet Archive et de retrouver celles qui correspondaient aux données de l'ADS, ainsi

[137] DST4L Class Notes - Google Docs, [sans date]. [en ligne]. [Consulté le 26 juillet 2014]. Disponible à l'adresse : https://docs.google.com/document/d/1WUz4UwwRv5szcsODIwcEV7qAGNc0gjL-oDErFQ2MoBY/edit?pli=1

[138] ERDMANN. 2014. p.24. « Another goal of DST4L was to upgrade the skills of librarians, and many of the participants are now using their new-found skills. For instance, Veronica Downey has automated library processes using Python, Alex Holachek is helping the NASA ADS improve its visualization tools, and Katie Frey is implementing technologies in astronomy ».

[139] Nous traduisons ici le titre des deux posts du blog de la DST4L qui nous serviront d'exemple ici : How to Beat Bibliographic Data into Submission, pt. 1 | Data Scientist Training for Librarians, [sans date]. [en ligne]. [Consulté le 7 juillet 2014]. Disponible à l'adresse : http://altbibl.io/dst4l/how-to-beat-bibliographic-data-into-submission-pt-1/ et How to Beat Bibliographic Data into Submission, pt. 2 | Data Scientist Training for Librarians, [sans date]. [en ligne]. [Consulté le 7 juillet 2014]. Disponible à l'adresse : http://altbibl.io/dst4l/how-to-beat-bibliographic-data-into-submission-pt-2/

[140] *Ibid*., pt. 1.

que celles qui n'y correspondaient pas, révélant ainsi les manques qui peuvent exister dans les collections numériques de la NASA. Grâce à ce procédé, les participants ont pu construire un tableur[141] dans lequel figurent, d'une part, les données récupérées de l'ADS et d'autre part celles récupérées de l'Internet Archive, chaque ligne du tableau faisant correspondre (ou non) les données de l'ADS et de l'IA. De simples tris croisés permettent ensuite de repérer les doublons et les manques d'une collection à l'autre.

Les participants au projet ont ensuite cherché à visualiser la totalité des données récupérées de l'Internet Archive dans le domaine de l'astronomie[142] et ont généré ces visualisations en fonction des questions qu'ils avaient à poser à leurs données. Une première visualisation sous forme de pastilles colorées devait nous renseigner, par exemple, sur les ouvrages anciens les plus téléchargés en astronomie ; un diagramme en barre nous indiquant ensuite quelle bibliothèque d'astronomie ayant participé à la collection numérique de l'Internet Archive possède la collection la plus précieuse. Au moyen d'une carte, les participants ont également choisi de représenter les lieux de publications les plus actifs dans le monde, toujours dans le domaine de l'astronomie. De même, une carte arborescente est chargée de représenter les ouvrages dont la numérisation a été la plus coûteuse. Enfin, un nuage de tags permet de visualiser les langues les plus courantes dans lesquelles sont publiés ces ouvrages d'astronomie, l'anglais et le français étant les deux langues les plus courantes.

Contrairement à ce que l'on pourrait penser au vu de ces réalisations, les techniques utilisées sont accessibles à des personnes qui n'ont pas nécessairement de bagage en informatique ou en design. En effet, si l'on observe pour commencer la manière dont les données ont été collectées, il faut avoir à l'esprit que l'Internet Archive a ouvert ses données et mis à la disposition de tous un formulaire de recherche qui permet de les récupérer aisément[143], et il en va de même pour l'ADS[144], puisque la NASA, du fait de son statut d'agence gouvernementale, a été dans l'obligation d'ouvrir ses données. Quant aux formats des données fournies par ces sources, il s'agit de XML pour l'ADS d'une part, et de JSON pour l'IA d'autre part, deux formats donc très lisibles pour des ordinateurs et permettant facilement la réutilisation des données, contrairement à des fichiers PDF, Word et JPEG qui peuvent être affichés mais non lus par un ordinateur. L'absence de droit de propriété intellectuelle sur les données de l'ADS et de l'IA permettent par ailleurs de les réutiliser librement. En ce qui concerne la structuration des données, le logiciel OpenRefine a également permis de se passer de l'écriture d'un code notamment pour faire passer les données d'un format à un autre. Par ailleurs, OpenRefine s'avère être un outil facile d'utilisation et utile à l'apprentissage de la programmation[145] .

De même, l'outil de visualisation utilisé, Tableau Public, est un outil prêt à l'emploi plutôt qu'une visualisation faite sur mesure[146] grâce au code : Tableau

[141] PRENTICE, Jennfer, ALSTINE, Colin Van, BENSON, Amy et FORD, Jacqueline, 2013. *ADS Monograph Matches in the Internet Archive (Excel)* [en ligne]. juin 2013. [Consulté le 19 août 2014]. Disponible à l'adresse : http://figshare.com/articles/ADS_Monograph_Matches_in_the_Internet_Archive/710921

[142] How to Beat Bibliographic Data into Submission, pt. 2.

[143] Internet Archive Search Engine. [Sans date]. Consulté le 19 août 2014. Disponible à l'adresse Web : http://archive.org/advancedsearch.php#raw.

[144] SAO/NASA ADS Custom Query Form, [sans date]. [en ligne]. [Consulté le 19 août 2014]. Disponible à l'adresse : http://adsabs.harvard.edu/abstract_service.html

[145] ERDMANN, 2014. « OpenRefine is a helpful stepping stone to the more advanced training in Python. The OpenRefine interface allows you to run simple funcions and regular expressions while hiding some of the complexities of programming. It also allows you to perform some data analysis ». p. 23.

[146] Cette distinction que nous faisons entre visualisation « prête à l'emploi » et « sur mesure » provient de l'ouvrage de Nathan Yau : « Certains logiciels, de type glisser-déplacer, sont prêts à l'emploi. D'autres nécessitent un peu de programmation. Cependant, il existe aussi des outils qui n'ont pas été conçus spécifiquement pour les graphiques de

permet en effet de charger les données sur son serveur puis de créer un affichage interactif des données et de publier ce dernier sur un site web ou un blog, comme l'ont fait les apprentis « donnéethécaires » de la DST4L. Le fait que ce logiciel ne requière pas de connaissances en programmation est mentionné parmi les raisons invoquées pour justifier son emploi[147].

On peut donc constater que ce travail effectué à partir des données, s'il n'est pas sans difficulté, reste à la portée des compétences d'un bibliothécaire. Par ailleurs, l'emploi de ces outils permet également l'apprentissage de langages de programmation qui peuvent permettre de ne pas se contenter d'une simple comparaison entre deux collections : on aura noté en effet que le travail présenté sur ce blog ne témoigne pas d'analyses statistiques très poussées, mais constitue une première étape vers ces dernières. Les visualisations présentées sont néanmoins déjà fort utiles pour leur qualités communicationnelles.

L'APPORT DE LA VISUALISATION POUR LA COMMUNICATION.

Dans son mémoire sur les relations entre le directeur de la bibliothèque et ses tutelles administratives et politiques[148], Marie Baudière explique notamment que les données des bibliothèques sont au cœur de la communication du bibliothécaire en direction de son élu :

> « Les directeurs de bibliothèque cherchent donc comment présenter à leurs élus l'activité de la bibliothèque. Tous attribuent au bilan annuel cette fonction : les données statistiques qui y figurent, les analyses sur la politique documentaire, la programmation culturelle, l'avancée des projets leur semblent à même de donner à l'élu une image réelle du fonctionnement de la médiathèque. (…) Un autre [conservateur] le conçoit comme un outil de découverte de la bibliothèque pour l'élu : "Il faut arriver à faire découvrir des choses aux élus. Les décideurs ne connaissent pas les métiers précis, ils ont une idée sur la bibliothèque mais il faut arriver à leur faire comprendre l'activité avec des données chiffrées. Attention, il faut choisir les informations les plus frappantes, les bons chiffres." Pourtant, la plupart des directeurs reconnaît que le bilan annuel est généralement un document trop complet, trop complexe que l'élu ne lit pas »[149].

Un double constat semble donc émerger de cette étude : d'une part, celui de l'efficacité des données pour expliquer à l'élu l'activité du service de lecture publique de sa collectivité et, d'autre part, celui de la trop grande complexité des bilans annuels où sont habituellement présentées les données statistiques qui concernent la bibliothèque. Or, la visualisation des données possède un certain nombre de qualités intrinsèques qui permettent de présenter de manière plus efficace l'information. Les bibliothécaires américains l'ont bien compris, eux qui ont multiplié les posts de blogs s'appuyant sur la visualisation. Citons le blog de l'OCLC, hanging together, sur la question des PEB par exemple[150], ou encore celui entièrement consacré à la visualisation des données des bibliothèques publiques, intitulé « Visualisation des données des bibliothèques : utiliser

données, mais qui se révèlent néanmoins utiles. Le présent chapitre traite de ces différentes options ». YAU, Nathan, 2013. *Data visualisation: De l'extraction des données à leur représentation graphique*. Editions Eyrolles. p. 65.

[147] How to Beat Bibliographic Data into Submission, pt. 2. « Our group chose to work with Tableau for four main reasons :

1) Stellar visualizations ! So pretty !

2) You can work with multiple data sources simultaneously

3) It has a large visualization toolset and suit of graphics to choose from.

4) Doesn't require a background in coding ».

[148] BAUDIÈRE, Marie, 2013. *Le bibliothécaire, son élu, son directeur Marie Baudière.* Bibliothèque numérique de l'Enssib. Consulté le 20 août 2014. Disponible à l'adresse Web : http://www.enssib.fr/bibliotheque-numerique/documents/64142-le-bibliothecaire-son-elu-son-directeur.pdf.

[149] *Ibid*. p. 61.

les statistiques des bibliothèques publiques américaines »[151]. Plus proche de nous, le blog français « Bibliothèque [reloaded] » a consacré quelques pages, sous la plume d'Étienne Cavalié, à une expérience de visualisation de données grâce à l'outil Gephi[152] : l'auteur s'est en effet proposé de cartographier le réseau documentaire de son SCD (figure ci-dessous[153]).

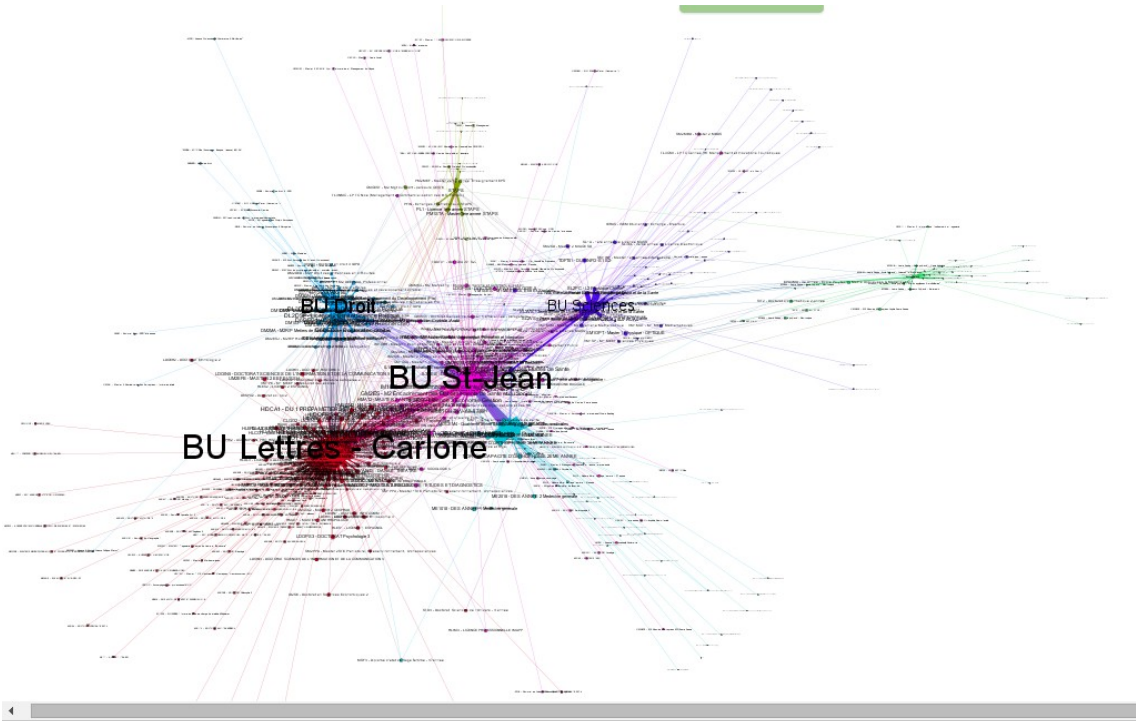


Figure 4 : Visualisation d'un réseau de SICD par Étienne Cavalier

La visualisation des données connaît donc un succès grandissant, notamment dans le milieu des bibliothèques. Nous nous proposons dès lors de tenter de donner quelques éléments d'explication à ce succès.

Séduire...

Comme l'écrit Marie Baudière dans son mémoire, « pour le directeur de la bibliothèque, le principal objectif de ses contacts avec son élu est le convaincre » :

> « Les stratégies de communication qu'il déploie pour cela sont multiples car l'asymétrie de son rapport hiérarchique avec l'élu lui impose une plus grande créativité. (…) la question des stratégies de conviction, de la prise en compte de l'interlocuteur dans la construction de l'argumentaire au management de l'élu, se rapproche parfois de celle de la séduction ou de la manipulation.[154] »

[150] Visualizing Network Flows: Library Inter-lending | hangingtogether.org, [sans date]. [en ligne]. [Consulté le 3 juin 2014]. Disponible à l'adresse : http://hangingtogether.org/?p=3053

[151] Library Data Visualization, [sans date]. [en ligne]. [Consulté le 20 mai 2014]. Disponible à l'adresse : http://librarydatavisual.blogspot.fr/

[152] CAVALIÉ, Etienne, [sans date]. Mais que fait Gephi ? *Bibliothèques [reloaded]* [en ligne]. [Consulté le 17 juillet 2014]. Disponible à l'adresse : http://bibliotheques.wordpress.com/2014/07/03/mais-que-fait-gephi/

[153] grapheprc3aats.png (Image PNG, 1024 × 1024 pixels) [sans date]. [en ligne]. [Consulté le 20 août 2014]. Disponible à l'adresse : https://bibliotheques.files.wordpress.com/2014/07/grapheprc3aats.png.

[154] BAUDIERE. 2014. p. 53-54.

On voit ici que la communication du bibliothécaire en direction de son élu nécessite une certaine créativité de sa part afin d'en appeler aux sentiments de son interlocuteur. Or, selon Nathan Yau, la visualisation répond tout à fait à ces critères[155] : en effet, bien loin de présenter un simple résumé de données quantitatives, la visualisation fait appel au design. Les qualités esthétiques qu'elle met en œuvre doivent permettre de susciter une réaction, positive ou négative, de la part de l'observateur. Pour illustrer cette idée, Nathan Yau prend l'exemple de la visualisation intitulée *We Feel Fine*, de Jonathan Harris, qui doit permettre d'explorer les émotions humaines à l'échelle mondiale[156]. Plus proche des bibliothèque, l'exemple du « navigateur de thèse » de Stanford, qui permet d'embrasser du regard l'évolution temporelle de sujets des thèses soutenues dans cette université (figure ci-dessous[157]).

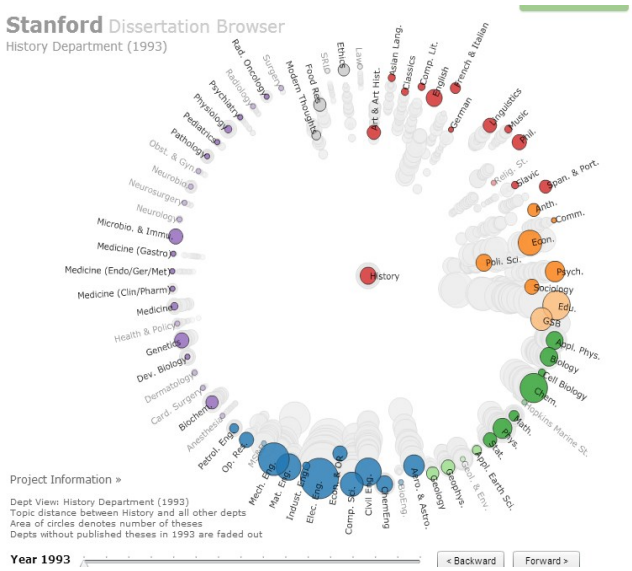


Figure 5 : Dissertation browser : l'Histoire en 1993 à Stanford.

Il nous semble en effet que la disposition circulaire, aérienne et colorée de chaque discipline est agréable à voir. D'un clic sur l'un des points, il est possible de voir se rapprocher ou s'éloigner les autres disciplines qui lui sont attenantes, ce qui permet de se faire assez rapidement une opinion sur l'évolution d'une discipline en particulier. L'histoire par exemple fait se rapprocher et s'éloigner d'elle successivement les sciences de l'éducation ou encore la sociologie : la taille du point étant fonction du nombre de thèses publiées dans la discipline, marquant ainsi l'évolution de l'intérêt des étudiants pour cette dernière[158]. Voilà qui permettrait, à ce qu'il nous semble, d'informer efficacement un président d'université sur l'activité scientifique de son institution.

Illustrer...

La visualisation permet également de résumer et de donner plus de clarté à un argumentaire : il s'agit clairement pour le bibliothécaire de donner à son interlocuteur une vision de son établissement. Nous pourrions ainsi prendre l'exemple d'une étude de cas développée dans l'article d'Anne C. Elguindi et de Bill Mayer intitulé « Raconter l'histoire de votre bibliothèque : comment tirer le meilleur parti de vos données dans une

[155] YAU, Nathan, 2013. *Op.cit*.

[156] HARRIS, Jonathan, KAMVAR, Sep. [sans date]. We Feel Fine. [en ligne]. [Consulté le 20 août 2014]. Disponible à l'adresse : http://wefeelfine.org/. *Cf* annexe p. 103, figure 19.

[157] Dissertation Browser | Information, [sans date]. [en ligne]. [Consulté le 23 mai 2014]. Disponible à l'adresse : http://www-nlp.stanford.edu/projects/dissertations/

[158] *Cf* annexe p. 104, figure 20 et 21.

présentation »[159], pour illustrer ce propos. En effet, des trois cas mentionnés de bibliothèques universitaires américaines, c'est le premier qui nous intéresse le plus, car la visualisation y est véritablement considérée selon sa dimension de représentation d'un objet.

En l'occurrence, il s'agissait d'illustrer un problème d'espace dans la bibliothèque, problème qui, d'ailleurs, est plutôt partagé par un grand nombre de bibliothèques universitaires. Afin de pouvoir offrir davantage d'espaces de travail aux étudiants, la direction de cette bibliothèque a pris la décision de déménager les revues imprimées dans un magasin distant. Il restait cependant à faire accepter cette décision à la tutelle de la bibliothèque en question, ce qui passait nécessairement par une prise de conscience de sa part de la situation de contrainte spatiale. « Étant donné que la décision de déménager les journaux imprimés étaient venue de l'analyse de quelques sources de données très divergentes (…), écrivent Elguindi et Mayer, il a paru préférable d'utiliser les données pour en dresser un tableau complet à la communauté universitaire »[160].

Afin de « dresser un tableau » de la situation, une première idée peut être de montrer à son interlocuteur des photographies des rayonnages surchargés de la bibliothèque. D'ailleurs, certains directeurs de bibliothèques, d'après Marie Baudière, n'hésitent pas à organiser des visites de la bibliothèque en direction de leurs élus afin qu'ils puissent se faire une image de leur établissement[161]. C'est là un moyen effectif mais qui ne prend pas en compte le fait qu'élus et directeurs n'ont pas toujours beaucoup de temps à consacrer à ces visites. Il peut donc s'avérer plus efficace de présenter les données de la bibliothèque visuellement. De simples diagrammes, pour commencer, peuvent faire l'affaire : deux courbes sur un même graphique, tel que celui présenté par Elguindi et Mayer[162], peuvent représenter le volume réel des rayons de la bibliothèque d'une part et le nombre de livres possédés par l'institution d'autre part : alors que la première variable reste stable et n'augmente plus, la seconde augmente constamment, ce qui met bien en évidence l'inéluctabilité de la saturation des espaces. Un second graphique[163] met en scène, sous la forme d'un diagramme en barres, le nombre d'étagères pleines d'une part et le nombre total d'étagères d'autre part , la barre correspondant à la première variable étant placée à l'intérieur de la seconde, ce qui, par superposition, permet d'observer le mince écart quantitatif des deux variables : c'est là une autre manière de représenter la saturation. Enfin, une dernière figure représente sous la forme d'une balance l'idée que les espaces de la bibliothèques consacrés à l'apprentissage empiètent nécessairement sur les espaces consacrés aux collections[164].

Ces figures sont bien sûr très simples et n'apprennent pas grand chose sur l'objet qu'elles doivent représenter, mais elles constituent une première image de l'activité de l'établissement et permettent d'ouvrir un dialogue entre la bibliothèque et sa tutelle.

[159] ELGUINDI, Anne C., MAYER, Bill. « Telling your library's story : how to make the most of your data in a presentation » dans ORCUTT, 2010. p. 26-28.

[160] *Ibid*. « As the decision to move out the bound journals had come from the analysis of some highly divergent sources of data (shelving statistics ; usage statistics of print and online journals and monographs ; computer use statistics ; physical plant statistics ; and an examination of what makes a library), it seemed best to use data to paint a full picture to the university community ». p. 26.

[161] BAUDIERE, 2014. p. 62.

[162] ELGUINDI, MAYER. 2010. Figure 3.2 p. 27.

[163] *Ibid*. Figure 3.3. p. 27.

[164] *Ibid*. Figure 3.4. p. 28.

Synthétiser...

Dans son mémoire, Marie Baudière met l'accent sur une des préoccupation des directeurs de bibliothèque en ce qui concerne leur communication en direction des élus, à savoir le fait d'aller au plus court et au plus parlant[165]. Du fait de son caractère d'immédiateté, la représentation visuelle répond bien à cette contrainte temporelle inhérente à la communication en direction des tutelles de la bibliothèque. Qui plus est, la visualisation des données a vocation, par définition, à être synthétique, comme l'écrit Lev Manovich, auteur d'un article intitulé « Qu'est-ce que la visualisation » ?[166] :

> « L'infovis utilise des éléments graphiques tels que des points, des lignes droites, des courbes et des formes géométriques simples afin de représenter les objets et leur relations entre eux, sans tenir compte de savoir s'il s'agit de personnes, de leurs relations sociales, des prix en bourse, des revenus nationaux, des chiffres du chômage, ou quoi que ce soit d'autre. (…) Cependant, le prix à payer de cette capacité est une extrême schématisation : nous rejetons 99% de la spécificité de chaque objet pour n'en représenter qu'1%, dans l'espoir que ces 1% nous révèlent des tendances parmi les caractéristiques de ces objets »[167].

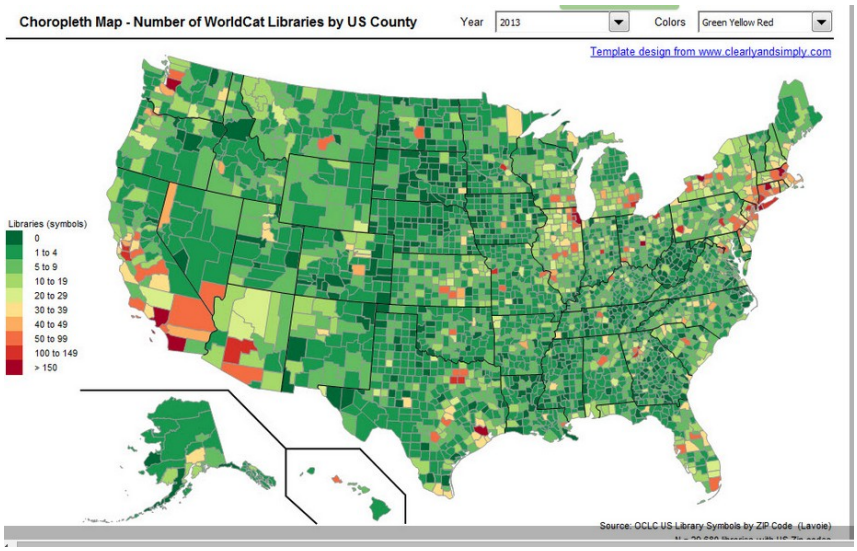


Figure 6 : Essai de représentation géographique de l'offre documentaire aux États-Unis : visualisation des bibliothèques sous forme de dégradé de couleurs

L'extrême capacité de schématisation et de simplification est donc une caractéristique propre à la visualisation des données : une bonne visualisation en effet, est celle qui a su éliminer tout ce qui paraissait superficiel par rapport à ce qu'elle cherche à montrer. En ce sens, elle est un excellent exercice de synthèse. Constance Malpas, chercheur à l'OCLC, démontre cela par les visualisations expérimentales qu'elles élabore à partir des données de WorldCat : dans un post du blog Hanging

[165] « L'élément "stratégique" le plus récurrent dans les réponses des directeurs de bibliothèque est l'élaboration de documents synthétiques car selon sa représentation sociale, l'élu est toujours pressé, il a peu de temps disponible. Ce point fait d'ailleurs l'objet de demandes spécifiques de la part des élus : "Il n'y a pas vraiment de qualités requises pour les documents que doit me remettre le directeur de bibliothèque, mais lors de son entretien, il a été porté attention sur le fait que les documents doivent être synthétiques" ». BAUDIERE, 2014. p. 54.

[166] Lev Manovich – What is Visualization? | Data Visualisation, [sans date]. [en ligne]. [Consulté le 30 juin 2014]. Disponible à l'adresse : http://www.datavisualisation.org/2010/11/lev-manovich-what-is-visualization/

[167] *Ibid*., p. 4. Ce caractère de schématisation extrême et de réduction est bien évidemment à rapprocher de la réduction opérée par les algorithmes : la visualisation, au même titre que ces derniers, est un média et en tant que tel, est dotée des mêmes limites épistémologiques. Néanmoins, le caractère schématique est peut-être plus évident dans la visualisation que dans les algorithmes, la procédure de ces derniers n'étant souvent pas connue pour des utilisateurs lambda.

Together[168], elle décrit en effet les étapes de sa recherche d'une représentation visuelle adaptée pour « modéliser l'offre et la demande à l'intérieur et à l'extérieur d'un consortium de bibliothèque, afin d'informer des décisions concernant la conservation locale et partagée de collections imprimées »[169].

L'information recherchée n'est donc pas des plus simples. Pour autant la visualisation cartographique permet de s'en faire une idée claire assez rapidement : après avoir renoncé à représenter les bibliothèques américaines sous forme de points[170], du fait de l'illisibilité que cela induisait, l'auteur s'est finalement rabattue sur une carte choroplèthe (figure ci-dessus[171]) dans laquelle les dégradés de couleur permettent de « montrer comment la demande est distribuée à une échelle "au-delà de l'institution" », afin de « comprendre le rôle de la logistique dans l'optimisation de la circulation des ressources des bibliothèques »[172].

Ces visualisations à grande échelle, telles que celles produites par l'OCLC à partir des données de WorldCat sont aussi très utiles pour comparer les bibliothèques entre elles.

Comparer...

> « (…) Tenir compte de l'image de l'élu est un des éléments des stratégies déployées par les directeurs de bibliothèque en agissant notamment sur l'émulation entre les collectivités ; un directeur de bibliothèque interrogé expliquait que pour réussir à obtenir un budget d'investissement important pour un projet qui lui semblait prioritaire, il s'était informé auprès du Ministère de la Culture et de la Communication et aussi auprès de collègues pour connaître l'avancement des autres équipements similaires sur ce type de chantier afin de situer sa propre bibliothèque. Il avait ensuite fait une note, qu'il avouait avoir un peu poussée, sur cette question en montrant le retard de sa bibliothèque »[173].

La comparaison paraît être un élément fondamentale de la communication des bibliothécaires en direction de leur tutelle institutionnelle. Or, la visualisation des données, lorsqu'elle est faite à grande échelle, permet de mettre en place ce type de comparaison entre établissement. Nous pourrions reprendre ainsi l'exemple de la visualisation développée au sein de la DST4L, notamment la première[174], destinée à répondre à la question de savoir quel ouvrage ancien d'astronomie avait été le plus téléchargé à partir de la bibliothèque numérique de l'Internet Archive. Cette visualisation est interactive : en cliquant dans le menu à gauche de l'image sur l'une des institutions ayant fourni des ouvrages à la bibliothèque, le fond de

[168] MALPAS, Constance. [sans date]. Sliding scale: mapping local, group and system-wide library infrastructure | hangingtogether.org. [en ligne]. [Consulté le 21 juillet 2014]. Disponible à l'adresse : http://hangingtogether.org/?p=3149

[169] *Ibid*. « My current objective is a lot more prosaic : modeling supply and demand within and outside of a given library consortium to inform decisions about local and shared stewardship of print collections ».

[170] *Top-250-CIC-borrowers-by-location.jpg (Image JPEG, 658 × 435 pixels)*, [sans date]. [en ligne]. [Consulté le 21 août 2014]. Disponible à l'adresse : http://hangingtogether.org/wp-content/uploads/2013/07/Top-250-CIC-borrowers-by-location.jpg. *Cf* annexe p. 105, figure 22.

[171] *Choropleth_US_libs_by_county.jpg (Image JPEG, 1017 × 653 pixels) - Redimensionnée (96%)*, [sans date]. [en ligne]. [Consulté le 21 août 2014]. Disponible à l'adresse : http://hangingtogether.org/wp-content/uploads/2013/07/Choropleth_US_libs_by_county.jpg. *Cf* annexe p. 105, figure 23.

[172] MALPAS. [sans date]. « For this, I think the county-level choropleth is actually quite useful. It helps to show how demand is distributed at 'above-the-institution' scale, and this is important for understanding the rôle of logistics in optimizing the flow of library ressources ».

[173] BAUDIERE, 2014. p. 55.

[174] Astronomy Texts in the Internet Archive, [sans date]. *Tableau Software* [en ligne]. [Consulté le 21 août 2014]. Disponible à l'adresse : http://public.tableausoftware.com/views/AstronomyTextsintheInternetArchive/Whatwasthetopdownloadedastronomywork?:showVizHome=no

l'image se grise, tandis que les pastilles de couleur correspondants à l'institution restent colorés, ce qui permet de les distinguer par rapport aux autres. De la sorte, on peut se faire une idée assez directe de l'importance de l'institution dans la collection globale de l'Internet Archive. Lorsque l'on clique sur la Fisher – University of Toronto, par exemple, on s'aperçoit qu'elle dispose des deux documents les plus téléchargés, mais qu'en dehors de ces documents, cette bibliothèque n'a pas fourni beaucoup d'autres ouvrages[175]. Lorsque l'on clique sur l'université d'Harvard, en revanche, on observe que les points rouges ne sont certes pas volumineux, mais nombreux, ce qui signifie que cette bibliothèque dispose de collections véritablement importantes en astronomie (figure ci-dessous[176]). Enfin, un clic sur la Duke University Library nous contraint à chercher du regard les rares petits points bleus lui correspondant sur l'image : de fait, l'institution n'a pas une collection ancienne très étendue dans le domaine de l'astronomie[177].

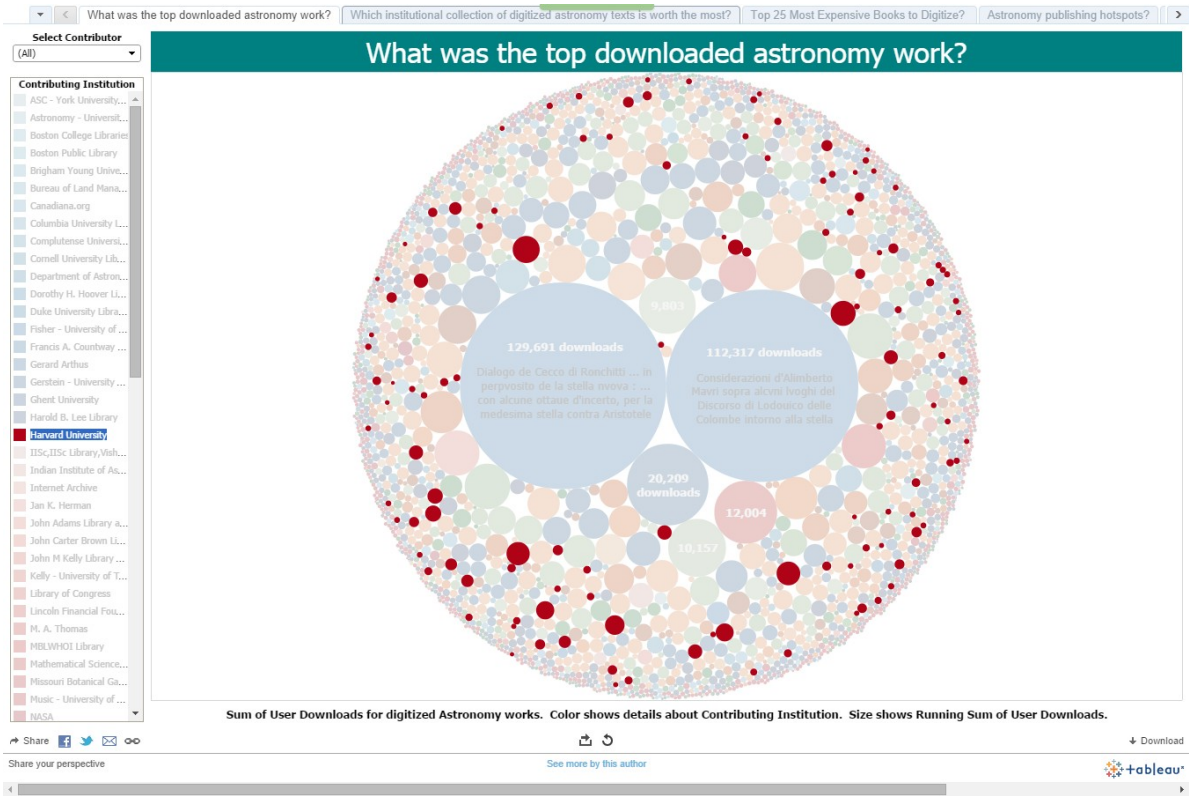


Figure 7 : La collection en Astronomie de la Bibliothèque d'Harvard

On pourra trouver ce type de comparaison dans un autre post de blog, celui de Dan Cohen, directeur exécutif de la DPLA[178] : les collections de chaque bibliothèque ayant participé à la DPLA ont été visualisées sous la forme de diagramme en barres, représentant le nombre d'ouvrages qu'elles possèdent en commun les unes avec les autres. Chaque diagramme permet de visualiser ce qui fait la particularité d'une bibliothèque et de sa politique d'acquisition : les diagrammes s'appuyant vers la gauche caractérisent des bibliothèques qui ont un grand nombre d'ouvrages qu'elles seules détiennent, comme Harvard par exemple, qui a mis l'accent sur la constitution d'une

[175] *Cf* annexe p. 106, figure 24.
[176] *Ibid*.
[177] *Cf* annexe p. 106, figure 25.
[178] COHEN, Dan, 2012. Visualizing the Uniqueness, and Conformity, of Libraries. *Dan Cohen* [en ligne]. 13 décembre 2012. [Consulté le 11 juin 2014]. Disponible à l'adresse : http://www.dancohen.org/2012/12/13/visualizing-the-uniqueness-and-conformity-of-libraries/

collection de livres rares[179], tandis que d'autres bibliothèques universitaires, comme la Lafayette College ont préféré constituer des collections plutôt universelles et accessibles à tous[180].

Il est vrai que ces visualisations ne nous apprennent pas grand chose sur les bibliothèques qu'elles représentent, surtout lorsqu'on les compare à des outils comme l'Observatoire de la Bibliothèque dont la précision est sans doute plus grande. Néanmoins, elles sont remarquables pour leur capacité à situer de manière visuellement agréable, précise, directe et globale l'activité d'un établissement, notamment lorsque l'on est en situation de devoir la mettre en valeur au cours d'un court entretien avec un élu ou un directeur d'université.

De la politique documentaire à la navigation dans les collections...

Processus de sélection informé par les données d'usages, la Patron-Driven Acquisition (PDA) est peut-être l'exemple type du pilotage d'un établissement documentaire par les données et incarne d'une certaine manière un transfert de responsabilité des acquisitions du bibliothécaire vers les utilisateurs de la bibliothèque. Or, comme le souligne Finbar Galligan, chef d'équipe marketing et communications chez Swets[181], « là où le contenu n'est plus fournit à l'avance, la recherche et la découverte deviennent les parties les plus importantes du flux de travail »[182] : la PDA suppose donc l'existence préalable d'un moteur de recherche permettant de moissonner des corpus complets de documents. D'une certaine manière, on peut dire qu'il y a bien un transfert de la politique documentaire, mais ce transfert ne se fait pas tant vers les usagers de la bibliothèque que vers les algorithmes qui fournissent les contenus en fonction de classements qui transposent les choix et présupposés de départ de leurs concepteurs[183].

Ce caractère médiatique des algorithmes fournisseurs de contenus est amplifié par l'ajout de fonction de recommandations basées sur les recherches passées et les recherches similaires faites par d'autres utilisateurs. Comme le suggère Finbar Galligan, ces évolutions sont en effet à envisager dans le cadre de bibliothèques de plus en plus pilotées par ses usagers (patron-driven)[184].

[179] *Cf* annexe p. 107, figure 26.

[180] *Cf* annexe p. 107, figure 27.

[181] Établissement qui se définit lui-même comme un gestionnaire de contenus électroniques en direction des bibliothèques et de leurs lecteurs. Content Management Services for Libraries and Publishers, [sans date]. [en ligne]. [Consulté le 8 décembre 2014]. Disponible à l'adresse : http://www.swets.fr/

[182] GULLIGAN, Finbar. Sans date. Patron-driven library - Patron-driven acquisition - Research Information. [en ligne]. [Consulté le 3 décembre 2014]. Disponible à l'adresse : http://www.researchinformation.info/features/feature.php?feature_id=485. « Where content isn't provided in advance, search and discovery becomes the most important part of the workflow. If the end-users can't find particular content unless they already know it exists, then the system will automatically fail ».

[183] Outre les théories de Neal Thomas que nous avons déjà citées en première partie (p. 24-26), les explications de Rachel Schutt et Cathy O'Neil viennent illustrer de manière pertinente ce propos : « Another way in which the assumption that N = ALL can matter is that it often gets translated into the idea that data is objective. It is wrong to believe either that data is objective or that "data speaks," and beware of people who say otherwise. We were recently reminded of it in a terrifying way by this New York Times article on Big Data and recruiter hiring practices. At one point, a data scientist is quoted as saying, "Let's pu everything in and let the data speak for itself.". If you read the whole article, you'll learn that this algorithm tries to find "diamond in the rough" types of people to hire. A worthy effort, but one that you have to think through. Say you decided to compare women and men with the exact same qualifications that have been hired in the past, but then, looking into what happened next you learn that those women have tended to leave more often, get promoted less often, and give more negative feedback on their environments when compared to the men. Your model might be likely to hire the man over the woman next time the two similar candidates showed up, rather than looking into the possibility that the company doesn't treat female employees well ». O'NEIL et SCHUTT. 2013. Non paginé dans sa version électronique.

[184] GULLIGAN, Finbar. Sans date. « Where search tools act as gatekeepers for nearly all scholarly content, they will need to be refined in not only the extent of their indexes, but also in the underlying algorithms that allow them to harvest, index and connect the wealth of content available across the net. (…). Advanced semantic techniques can aid the

Cependant, d'après Ronald E. Day cela aurait tendance à limiter, d'une certaine manière, le champ des découvertes possibles de contenus :

> « Les indices computationnels construits de manière récursive (de même que le classement algorithmique) peuvent réduire les potentialités intentionnelles des Mois aux possibilités logiques de personnes socialement reconnues à travers le renforcement des recherches précédentes et des recherches des autres (…). Ce qui dans l'analyse de citation commence comme des explications comportementales dans le but de l'analyse de citation se termine en algorithmes qui contrôlent la construction de l'identité ainsi que l'intention dans la recherche et la communication d'information à travers les présupposés psychologiques et sociologiques de groupes »[185].

De fait, l'affirmation selon laquelle la PDA serait au plus près des besoins réels des lecteurs en matière de documentation, contrairement à la traditionnelle politique documentaire impulsée par les spéculations des bibliothécaires concernant ces mêmes besoins, serait ainsi relativement erronée [186]: la PDA ne fait que remplacer la subjectivité des responsables de collections par celle des algorithmes et des usagers qui intériorisent eux-mêmes les présupposés bibliométriques de ces derniers.

Ce processus de médiatisation de la recherche documentaire, – par médiatisation est entendu ici la transformation d'un objet en média, c'est à dire en interface transposant l'opinion, juste ou non, d'un groupe social –, deviendrait particulièrement problématique à partir du moment où les algorithmes et leur produits seraient utilisés pour piloter la politique général d'un établissement scientifique, comme le suggère notamment Finbar Galligan :

> « [Cet] instantané à haute résolution pourrait ensuite être utilisé pour une multitude de buts, notamment : affiner les objectifs institutionnels, élaborer des programmes d'enseignement fondés sur des sujets qui sont utilisés par la faculté d'aujourd'hui, déterminer des ressources qui sont applicables à un cours en particulier et prédéfinies en fonction des usages actuels ou des données de micro-acquisitions sur plusieurs unités temporelles pour ce même cours, et permettre à la bibliothèque de développer des services auxiliaires autour de l'offre principale de contenus, qui serait elle-même largement automatisée à travers la construction d'une collection par l'action collective de tous les usagers de la bibliothèque »[187].

---

discovery process, linking individual pieces of content and making suggestions and connections that are relevant to a single researcher's profile and reading preferences. (…) Data at the microlevel of the simple researcher is interesting but it starts to become really useful when you can aggregate it up over several layers of granularity. This could mean that content will not only be recommended based on the individual preferences, but on similar researchers and what they are using, giving an automatic recommendation engine that can be scaled all the way up to institutional level ». Non paginé dans sa version électronique.

[185] DAY. 2014. « Recursively constructed computational indices (as well as algorithmic ranking can narrow the intentional potentialities of selves to the logical possibilities of socially recognized persons through the strenghtening of previous searches and the searches of others. (…) What in citation analysis start as behavioral explanations for the purpose of citation analysis end up as algorithms that control identity construction and intention in information searching and communication through group psychological and sociological assumptions ». p. 69.

[186] Affirmation que l'on peut trouver notamment dans le blog the ScholarlyKitchen : RICK ANDERSON, [sans date]. What Patron-Driven Acquisition (PDA) Does and Doesn't Mean: An FAQ. *The Scholarly Kitchen* [en ligne]. [Consulté le 6 décembre 2014]. Disponible à l'adresse : http://scholarlykitchen.sspnet.org/2011/05/31/what-patron-driven-acquisition-pda-does-and-doesnt-mean-an-faq/. Si, de fait, la « sagesse des foules » est un argument avancé pour justifier la préférence pour un modèle dirigé par les usages plutôt que par les bibliothèques, les présupposés et représentations véhiculées par les nouvelles techniques accompagnant la PDA incitent à se poser la question de savoir jusqu'à quel point les foules peuvent être sages. *Cf* « Des bouquets aux acquisitions faites par les usagers, un nouvel équilibre à trouver 5/7 », [sans date]. [en ligne]. [Consulté le 8 décembre 2014]. Disponible à l'adresse : http://www.bibliobsession.net/2011/03/03/du-bouquet-aux-acquisitions-faites-par-les-usagers-un-equilibre-a-trouver/

[187] GULLIGAN. Sans date. « The high-level snapshot could then be used for a host of purposes, including : refining institutional objectives ; building teaching programmes based on topics that are being used by the actual faculty ; determining preset materials that are applicable to a particular course, based on actual usage or micro-acquisition data over time for the same course ; and allowing the library to develop ancillary services around the core content offering, which would be largely automated based on the collective collection building of all library users ».

Si l'on reprend en effet la vision de Larry Page sous-tendant l'algorithme de PageRank dont il est le concepteur, ce dernier fonctionne essentiellement sur l'autorité des liens hypertextes qui, selon ses mots, « encodent une somme considérable de jugements humains latents (...) »[188]. Ainsi, prendre des décisions institutionnelles en se fondant sur les résultats d'un tel algorithme reviendrait finalement à se fier à des représentations des objets décrits par ces algorithmes plutôt qu'à une connaissance à proprement parler de ces objets : « étant donné les exemples passés de classements populaires ou 'de masses', l'application d'algorithmes enracinés dans la psychologie de groupe à la production de connaissance conduit à se demander à quoi la délégation de la connaissance à l'opinion en tant que telle pourrait ressembler »[189].

Si donc nous en venons à l'avenir à faire de la navigation virtuelle dans les collections de la bibliothèque le moteur de la gestion de ces mêmes collections et de l'établissement qui les fournit ou les contient, il est certain qu'il devient nécessaire de réfléchir à un moyen de mettre en avant la subjectivité inhérente à un tel système, subjectivité que les discours actuels autour de ces innovations tendrait en effet à occulter. C'est à un tel moyen que nous nous emploieront à consacrer le troisième temps de notre réflexion.

[188] Larry Page cité dans CARDON, Dominique, 2013. Dans l'esprit du PageRank. *Réseaux*. 1 avril 2013. Vol. 177, n° 1, pp. 63-95. DOI 10.3917/res.177.0063. p. 71.

[189] DAY. 2014. « Within the citation rat race and citation mongering, it becomes unclear what the rôle of truth is or how one can find a position for critique that itself is not a commodity or at least seen as a commodity and self-commodification. Instead, given past examples of popular or « mass » rankings, the expansion of algorithms rooted in group psychology to the production of knowledge lead one to wonder what the delegation of knowledge to opinion as well, may look like ». p. 73.

LES DONNÉES, UN OUTIL DE NAVIGATION DANS LES COLLECTIONS ?

Par leur caractère ontologique, les métadonnées ont un statut particulier par rapport aux autres types de données des bibliothèques : elles ont valeur de symbole par rapport aux objets de la collection qu'elles désignent. De ce fait, l'enjeu de faire parler les données des bibliothèques, lorsque ces données sont en réalité des métadonnées, est radicalement différent et peut-être plus important : il est par exemple désormais possible de construire des moteurs de recherche et des systèmes de recommandation taillés à la mesure de chaque usagers qui les utiliseraient, avec toutes les limites que nous avons déjà pu souligner. Ainsi, là où le caractère unique et standardisé de la classification permettait, dans le monde physique, de se déplacer dans les collections tout en visualisant une géographie du savoir, dans le monde numérique, une classification qui permet une navigation efficace de l'utilisateur est une classification qui s'adapte étroitement à la personnalité de l'individu : classification et navigation tendent alors à se confondre. Cette classification « sur mesure » aurait été impensable dans le monde physique, du fait des contraintes spatiales que ce dernier implique. Dans le monde numérique, c'est la multiplication des données, la « datafication » de notre environnement – y compris livresque –, qui permet un tel tour de force : les algorithmes classent à la fois les données personnelles d'un utilisateur et les données produites sur un objet pour proposer à ce dernier un ensemble de produits dont on suppose qu'il y portera intérêt[190]. À cela s'ajoute les performances de la fouille de texte[191] (ou text mining) : il s'agit d'un ensemble de techniques de linguistique, de statistique et d'apprentissage automatique visant à modéliser et structurer l'information contenue dans des ressources textuelles, ce, par exemple pour indexer un ensemble de documents et les classer selon leurs thèmes[192]. La « datafication » va donc jusqu'aux mots d'un texte pris comme unité et dans une certaine mesure transformé en métadonnées par le biais de l'analyse de contenu.

Néanmoins, que signifie dans ce contexte, la notion d'exploration des collections, quand « la découverte accidentelle » d'un objet reste « difficile avec le système de type requête-réponse utilisé pour les moteurs de recherche »[193] ?

> « Les concepteurs des services grand public comme Google Books et Amazon en sont conscients et ont mis en place plusieurs techniques alternatives, essentiellement basées sur la visualisation : des lectures en cours d'autres usagers, ou de recommandations inspirées de la navigation passée, voir de notices d'ouvrages sélectionnées de manière aléatoire. Dans tous les cas, il faut substituer au texte descriptif (notice de l'ouvrage) des indices visuels qui permettront une lecture de survol de l'ensemble de l'écran »[194].

Pour être honnête, le passage des nouvelles classifications offertes par le numériques à la navigation dans des collections virtuelles par le moyen de la visualisation des données est un domaine en voie d'expérimentation et nécessite de

[190] Nous renvoyons ici au deuxième chapitre de notre première partie, portant sur l'algorithme FRBR, comparé à ceux de Google et d'Amazon. p. 30.

[191] Fouille de textes, 2014. *Wikipédia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Fouille_de_textes&oldid=107660108. Page Version ID: 107660108

[192] Text mining, 2014. *Wikipedia, the free encyclopedia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://en.wikipedia.org/w/index.php?title=Text_mining&oldid=637280039. Version ID: 637280039

[193] CRAMER, Florian, CUBAUD, Pierre, DACOS, Marin, JAMES, Yannick, LANTENOIS, Annick (dir.). 2011 *Lire à l'écran : contribution du design aux pratiques et aux apprentissages des savoirs dans la culture numérique : [actes de la journée d'étude Lectures numériques, Valence, 11 mars 2010]*. Organisée par l'École supérieure d'art et design Grenoble-Valence. p. 57.

[194] *Ibid*.

« repenser la mise en espace de la bibliothèque numérisée, en développant des métaphores de navigations spécifiques »[195].

Nous nous proposons donc en dernier lieu d'observer dans quelle mesure la visualisation des collections est effectivement un atout pour la communication de la bibliothèque, mais aussi pour la navigation dans les collections : après avoir exposé plusieurs exemples de visualisations expérimentales à partir de la classification UDC, nous développerons des vues plus personnelles sur ce qu'il nous paraît intéressant d'envisager à l'avenir pour naviguer dans les collections à l'aide des données.

De la classification à la navigation...

Dans son essai intitulé *Tout est fragmenté*[196], David Weinberger développe l'idée que le bouleversement apporté par l'évolution vers le numérique ne réside pas tant dans la mutation de l'information en elle-même que dans l'accès à l'information à proprement parler. C'est ainsi que, pour le démontrer, il compare Amazon à Melvil Dewey :

> « En soi, Amazon est aussi éloigné que possible d'une bibliothèque appliquant la classification Dewey. Dewey a créé une manière unique de regrouper les livres : Amazon tâche d'en trouver autant que possible. Melvil Dewey s'est chargé lui-même de la conception du système : Amazon, quant à lui, laisse tout le monde créer ses propres catégories, leurs donner un nom amusant puis les publier. Dewey a privilégié la clarté et l'ordre, se prosternant devant les dieux de la métrique en créant un système basé sur des multiples de 10 : Amazon apprécie au contraire un désordre chaleureux, suggérant partout dans ses pages des manières alternatives de naviguer ainsi que des offres insolites particulières à chacun. Lorsque l'on cherche un livre dans une bibliothèque organisée sur le modèle de Dewey, on peut être très content de trouver un autre livre sur le même sujet juste à côté du premier sur l'étagère. Mais lorsque l'on cherche à acheter un livre sur Amazon, la sérendipité planifiée vous conduit vers un choix bien plus large de livres, déterminé par les éditeurs d'Amazon, les algorithmes ainsi que les autres consommateurs. Le système de Dewey privilégie la stabilité qui accompagne le monde physique – des livres sur des étagères, de l'encre blanche au dos des livres, tandis qu'Amazon se targue de sa capacité à grouper et regrouper de manière instantanée ses produits »[197].

Le propos de David Weinberger est donc d'affirmer que la transformation numérique, en créant des données à partir de toutes choses, est dans la capacité de nous faire découvrir une information bien plus importante quantitativement mais aussi qualitativement grâce à la diversité des propositions : les possibilités de trouver un objet inattendu seraient donc plus grandes que dans le monde physique. Pour l'auteur, ce bouleversement se caractérise par trois propriétés offertes par le

[195] *Ibid*. p. 59.

[196] WEINBERGER, 2008.

[197] *Ibid*. « Amazon itself is about as far from a Dewey-compliant library as one can get. Dewey created a single way to cluster books ; Amazon finds as many ways as it can. Melvil Dewey took the design of the system upon himself ; Amazon lets anyone create her own category, give it a fun name, and publish it. Dewey prized neatness and order, bowing to the metric gods when he created a system based on multiples of ten ; Amazon likes a friendly disorder, stuffing its pages with alternative ways of browsing and offbeat offers peculiar to each person's behavior. When you go to find a book in a Dewey-based library, you may be delighted to find another book on the same topic next to it on the shelf ; when you go to buy a book at Amazon, the planned serendipity shows you a far wider range of books, determined by Amazon's editors, algorithms, and fellow shoppers. Dewey's system prizes the stability that comes with the physical world – books on bookshelves, white ink on spines ; Amazon prides itself on its ability to cluster and recluster instantly ». p. 132.

numérique : d'abord, la remise en cause d'un système unique de classification de l'univers par le désordre numérique, ensuite, l'affranchissement des contraintes du monde physique grâce à la possibilité de créer des classification multidimensionnelles, s'adaptant instantanément selon les points de vue et permettant de disposer un même objet à plusieurs nœuds d'une classification, enfin, le passage d'une vision universelle de l'ordre de l'univers, perçue comme pouvant poser des problèmes de société, à une vision propre à chaque individu. Ainsi, le désordre numérique doit-il être, dans ce qu'envisage l'auteur pour les années à venir, à l'origine d'un changement radical de notre manière de percevoir le monde.

Peut-être serait-il bon d'exposer à la fois les bouleversements effectivement introduits par les données et les mythes qui, nous semble-t-il, ne manquent pas d'accompagner cette transformation.

« De l'Arbre au Labyrinthe »[198]

Dans son essai, David Weinberger fait remonter les origines de nos systèmes de classification actuels, qu'il s'agisse de la classification linéenne des espèces ou du système décimal de Dewey appliqué dans les bibliothèques, au premier système élaboré par Aristote et devenu au III^{e} siècle le célèbre arbre de Porphyre. Le principe de cet arbre est de regrouper l'univers dans un tronc commun, puis de diviser celui-ci en autant de branches qu'il y a de genres, eux-mêmes divisées en espèces puis en sous-espèces, et ainsi de suite jusqu'à l'individu : un chien est ainsi un exemplaire d'une race, appartenant elle-même à la sous-espèce des canidés, relevant quant à elle de l'espèce mammifère, cette dernière s'inscrivant au niveau supérieur dans le genre animal. La classification Dewey est elle-même organisée selon ce modèle puisque elle divise la connaissance en grands domaines appelés « classes » (regroupant philosophie, religion, sciences sociales, sciences de la nature et mathématiques, etc.), ces classes connaissant des divisions en disciplines et sous-disciplines, etc., jusqu'à en arriver à l'exemplaire unique du livre. Les livres, dans ce contexte, représentent les feuilles de l'arbre classificatoire. Or, tout le propos de David Weinberger consiste à dire que l'évolution vers le numérique tend à supprimer la métaphore unique de l'arbre pour ne garder que les feuilles, que l'on peut alors réorganiser à souhait selon ses propres catégories. Il donne l'exemple de la musique pour illustrer son propos :

> « Amazon veut nous vendre des livres. L'organisation qu'il donne a son offre n'est pas contrainte par une géographie sous-jacente. Amazon est capable de traiter son énorme collection de livres – à savoir les livres qu'il peut se procurer si quelqu'un en veut un exemplaire – comme un amas hétérogène qui peut être numériquement classé afin de refléter les intérêts individuels de chaque visiteur. (…) Le problème fondamental de Dewey ne réside pas dans le fait qu'il était un excentrique ou que sa première éducation était provinciale. Le véritable problème est que toute carte de la connaissance implique que la connaissance ait une géographie, qu'elle ait une vue surplombante, qu'elle ait une forme »[199].

Ainsi les acheteurs ont-ils accès directement aux feuilles que sont les livres, sans avoir à passer par les nœuds de l'arbre[200] que forme la classification Dewey. Ils

[198] Nous reprenons là le début du titre d'un recueil d'essais publié par Umberto Eco : ECO, Umberto, 2010. *De l'arbre au labyrinthe études historiques sur le signe et l'interprétation*. Paris : Grasset.

[199] WEINBERGER, 2008. « Amazon wants to sell us books. Its organization of its offering is not bound by underlying geography. Amazon is able to treat its enormous collection of books – that is, the books it can get if someone wants a copy – as a miscellaneous pile that can be digitally sorted to reflect the individual interests of each visitor. (…) This fundamental problem with Dewey's system is not that he was an eccentric or that his early education was provincial. The real problem is that any map of knowlegde assumes that knowledge has a geography, that it is a top-down view, that it has a shape ». p. 135.

[200] Ici rappelé sous la forme d'une carte, ce qui n'est pas incompatible, puisque une carte est formée sur le même principe qu'un arbre classificatoire

remplacent ensuite cet arbre unique par autant d'arbres classificatoires différents qu'il y a d'utilisateurs d'Amazon. Mais David Weinberger va plus loin encore dans l'explosion de l'arbre en feuilles, puisqu'il explique que la numérisation massive effectuée par Google a permis de faire d'une phrase, d'une expression ou même d'un mot une feuille qu'il est possible de brasser à l'infini avec d'autres éléments similaires afin de découvrir d'autres livres. Amazon produit ainsi, par le biais d'une fouille de texte[201], une analyse statistique du contenu d'un ouvrage, dont il rapproche les expressions les plus statistiquement significatives d'autres ouvrages employant des expressions similaires[202]. Le bouleversement numérique a donc fait explosé jusqu'au livre et à sa mise en page : la bibliothèque numérique va au-delà du livre.

Jusqu'à présent, il y aurait peu de chose à redire aux théories développées par Weinberger : sa vision de la stratégie des Amazon et autres géants du net semble juste. En revanche, il nous semble que la seconde partie de sa réflexion, portant sur les limites présupposées de la géographie sous-jacente à la connaissance, appelle une discussion. L'auteur explique en effet que la limite des systèmes classificatoires, tel que celui de Dewey, ne réside pas tant dans la vision du monde (ce à quoi l'auteur fait référence lorsqu'il parle de son excentricité ou de sa première éducation) que Dewey a fait transparaître dans la classification unique qu'il a proposé comme modèle à toutes les bibliothèques du globe, mais bien plutôt le fait que sa classification adopte une forme unique et invariable, peu adaptable aux désirs et aux goûts de chacun.

Tout d'abord, il convient de dire que même dans la nouvelle configuration numérique, la connaissance garde une géographie et une forme, celle du labyrinthe ou rhizome[203], plus couramment appelé « réseau ». En réalité, l'évolution actuelle de l'information n'est pas tant en voie de produire un bouleversement de notre manière de penser le monde, comme le pense David Weinberger, que l'inverse : c'est l'évolution de la pensée moderne qui, à ce qu'il nous semble, a fait naître les conditions nécessaires au bouleversement numérique actuelle et à notre nouvelle manière de rechercher et d'accéder à l'information. « Le système général des sciences et des arts est une espèce de labyrinthe, de chemin tortueux, où l'esprit s'engage sans trop connaître la route qu'il doit tenir »[204], écrit en effet d'Alembert dans le « Discours préliminaire » à l'*Encyclopédie*. D'Alembert exprime par là une idée chère aux lumières, à savoir le refus de « toute tentative de fonder un système a priori des idées » et la conception d'un savoir qui « s'articule comme une carte géographique sans frontières, sur laquelle des parcours infinis sont possibles »[205]. Le passage d'une classification unique à un « désordre » (selon le terme du titre de l'ouvrage de Weinberger[206]) que chacun parcours selon sa propre conception du monde avait donc déjà été entériné au temps des Lumières, ce que Weinberger est

[201] *Cf* note n°190.

[202] WEINBERGER. 2008. « For *The Little House Cookbook*, the list of "Statisically Interesting Phrases" includes "sterilizing kettle", "pie paste," "pastry surface," "buttered pie pan," and "blood-warm water". Click on any of these phrases and Amazon will show you other books that also use them : "sterilizing kettle" turns out to occur in *The Fall : A Novel*, by Simon Mawer ». p. 129.

[203] Cette pensée est développée par Eco, mais est également au cœur du projet de Gilles Deleuze et Félix Guattari intitulé « Capitalisme et Schizophrénie » : « Rhizome is a philosophical concept developed by Gilles Deleuze and Félix Guattari in their Capitalism and Schizophrenia (1972–1980) project. It is what Deleuze calls an « image of thought », based on the botanical rhizome, that apprehends multiplicities. » Rhizome (philosophy), 2014. *Wikipedia, the free encyclopedia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://en.wikipedia.org/w/index.php?title=Rhizome_(philosophy)&oldid=637871872. Page Version ID: 637871872

[204] ALEMBERT, Jean Le Rond d' et CONDORCET, Jean-Antoine-Nicolas de Caritat marquis de, 1821. *Œuvres de d'Alembert*. A. Belin. Volume 1, p. 44.

[205] ECO, 2010. p. 70.

[206] WEINBERGER, 2008. *Everything Is Miscellaneous: The Power of the New Digital Disorder*. *Op.cit.*

d'ailleurs prêt à concéder[207]. Cependant, il ne semble pas remarquer cette nouvelle géographie du savoir héritée de leur tradition, à savoir le labyrinthe en rhizome mis en lumière par Umberto Eco : « un modèle en réseau prévoit la définition de chaque concept (représenté par un terme) grâce à l'interconnexion de tous les autres concepts qui l'interprètent, chacun se tenant prêt à devenir le concept interprété par tous les autres »[208]. L'encyclopédie est bien fondée sur ce système, notamment à travers le système de l'index, qui permet le renvoi d'une notion à une autre indépendamment de leur classement alphabétique. Or, Eco souligne que c'est précisément cette conception du savoir en réseau qui est à l'origine des « ontologies » utilisées dans l'Intelligence Artificielle et, par extension, dans les nouvelles technologies numériques[209]. Par là peut-on dire que ce ne serait pas ces nouvelles technologies qui sont à l'origine de notre conception labyrinthique du savoir, mais plutôt l'inverse ? À l'image du Web sur le modèle duquel elles ont été conçues, les bibliothèques numériques ainsi que les catalogues en ligne sont donc formés en vastes réseaux labyrinthiques dans lesquels seules les multiples connexions entre les métadonnées permettent de se déplacer d'une information à une autre : c'est précisément pour pouvoir se déplacer plus facilement dans ce réseau que la tendance est en ce moment à l'ouverture des catalogues au web sémantique[210].

Enfin, là où David Weinberger tente d'expliquer que les nouvelles technologies utilisées par Amazon n'imposent désormais plus une vision unique de l'ordre des connaissances, mais permettent au contraire à chacun de se constituer librement sa propre classification, nous pourrions objecter que ce type de proposition ignore délibérément les présupposés inhérents aux algorithmes utilisés par Google et Amazon, auxquels nous avons déjà fait allusion dans la première partie de cette étude. En somme, nous pourrions dire que les géants du net, dans leur volonté de faire parler les métadonnées grâces aux nouvelles technologies offertes par l'ère du numérique, n'ont fait que déplacer la subjectivité et le caractère monopolistique propre à la classification traditionnelle au champ de la navigation quotidienne que nous effectuons dans leur corpus de données.

De l'universalité de la classification à l'individualité de la navigation

Une autre des théories développées par David Weinberger consiste à dire que le numérique nous ferait passer d'une classification unidimensionnelle, par nécessité physique, à une classification multidimensionnelle, caractérisée par le fait que l'on puisse placer un même objet à plusieurs endroits différents de la classification. Il illustre son idée de cette manière :

> « Mettons que vous vouliez un exemplaire du *Livre de cuisine de la Petite Maison dans la Prairie : la Cuisine de la Frontière d'après le classique de Laura Ingalls*, de Barbara Walker. Si vous cherchez le titre à la Bibliothèque Publique de New York, vous trouverez cinquante-deux exemplaires répartis dans ses nombreuses annexes. La plupart le range dans la section jeunesse.
> (…) Si vous souhaitez voir tous les livres portant à la fois sur la cuisine et sur l'histoire, sans spécifier que ces livres doivent être pour enfants ou être associés à un classique de la littérature, Amazon construira avec joie cette liste pour vous. C'est

[207] Il évoque effectivement mais de manière assez brève, les encyclopédistes. WEINBERGER, 2008. p. 25.

[208] ECO, 2010. p. 79.

[209] *Ibid*. p. 82-83. « Dans le cadre des recherches les plus récentes de l'Intelligence artificielle et des sciences cognitives, le thème des réseaux sémantiques a donné naissance à une théorie des ontologies. En dépit de son utilisation impropre, ce concept d'"ontologie", dont la signification philosophique est toute autre, désigne l'organisation catégoriale d'une portion d'univers qui prend la forme de n'importe quel arbre classificatoire ou réseau sémantique ».

[210] Le web sémantique est lui-même fondé sur les principes de l'ontologie, et forme un « réseau sémantique » : Ontologie (informatique), 2014. *Wikipédia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Ontologie_(informatique)&oldid=109058774. Page Version ID: 109058774

comme si l'on avait un système de classification décimal de Dewey écrit sur commande »[211].

L'auteur oppose donc la flexibilité de la classification numérique à l'uniformité de la classification élaborée dans le monde physique, l'innovation technologique étant tenue responsable du passage de l'un à l'autre. Pourtant, Umberto Eco démontre que l'arbre classificatoire hérité d'Aristote possédait une certaine souplesse, ce dernier ayant finalement renoncé à construire un arbre unique pour construire plusieurs classifications correspondant à des objectifs chaque fois différents[212].

Ces réflexions laissent à penser que ce ne serait pas tant en raison des contraintes physique ou des nouvelles opportunités offertes par le numérique[213] qu'en raison de choix politiques, que nous aurions abandonné le système classificatoire unique seulement au début du XXI^{e} siècle. Les Lumières, en effet, ont remis en cause l'arbre classificatoire unique pour lui substituer une vision labyrinthique de la connaissance, mais il ne s'agit pas que de cela : l'Encyclopédie est devenue elle-même une méthode d'enquête « à travers la bibliothèque générale et omnivore de la culture toute entière »[214]. Cette manière multidimensionnelle et labyrinthique de rechercher la connaissance et d'y accéder qui serait la nôtre aujourd'hui ne daterait donc pas des changements introduits récemment par les technologies numériques[215]. Nous pourrions plutôt envisager ces dernières comme le produit d'une évolution philosophique déjà vieille de quelques siècles.

De fait, en soutenant la thèse d'un bouleversement de nos modes d'accès à la connaissance introduit par le numérique, et non résultat d'une évolution antérieure de la pensée, il semble que David Weinberger cède aux sirènes de la neutralité et de l'objectivité dont nous avons déjà souligné, dans notre premier chapitre, l'impossibilité : selon son point de vue, un établissement public tel que la bibliothèque se devrait d'être laïc jusque dans sa classification, en ne privilégiant pas outre mesure une philosophie ou une religion par rapport à une autre[216]. À

[211] WEINBERGER, 2008. « Let's say you want a copy of the Little House Cookbook : Frontier Food from Laura Ingalls Wilder's Classic Stories, by Barbara M. Walker. If you look up the title at the New York Public Library, you'll find fifty two copies across the many branches. Most put it in the children's room, but the Donnell Library puts it in the reading room. Everyone of those branches, however, has it listed under its call number : 641.59 W. That translates to : Technology and applied sciences > Home economics and family living > food and drink. That's one logical place for it. But just one. If you search for the same book at Amazon, you'll find a similar classification scheme. But Amazon lists *The Little House Cookbook* under three categories :

_ Children's Books > Author & Illustrators, A-Z > (W) > Williams, Garth.

_ Children's Books > History & Historical Fiction > United States > 1800's.

_ Children's Books > Sports & Activities > Cooking.

(…) If you want to see all books about both cooking and history without specifying that the books have to be associated with a work of literature, Amazon will happily build that list for you. It's like having a Dewey Decimal Classification System written to order ». p. 126.

[212] « Porphyre trace un arbre des substances unique, tandis qu'Aristote utilise la méthode de la division avec beaucoup de précaution, voire de scepticisme. (…) En théorie, nous sommes autorisé à avancer l'hypothèse qu'il [Aristote] n'aurait pas su construire un arbre de Porphyre fini, et même en pratique (…), car nous le voyons dans *Les parties des animaux*, renoncer de fait à construire un arbre unique et réajuster des arbres complémentaires au gré de la propriété dont il veut expliquer la cause et la nature essentielle (… ». ECO, 2010. p. 20.

[213] Nous avons déjà vu que c'est plutôt la pensée qui a précédé ces dernières.

[214] *Ibid*., p. 72.

[215] Sur cette question, lire WRIGHT, Alex, 2008. *Glut: Mastering Information Through the Ages*. Cornell University Press.

[216] D'où peut-être cette réflexion que l'on peut lire dans son essai : « Comme l'écrit Wayne A. Wiegand, le biographe de Dewey, l'organisation de la connaissance qu'a produite Dewey a matérialisée "une vision du monde et une structure de la connaissance enseignée sur le campus de l'université d'Amherst entre 1870 et 1875" – une vision du monde et une structure qui présupposait que l'Occident était la culture la plus avancée et que le Christianisme était au fondement de la vérité » *Ibid*. p. 116. Il s'agit donc là d'une tradition libérale, qui prône le fait que la puissance publique doive s'abstenir de se prononcer sur ce que peuvent être les valeurs d'une vie bonne. L'humanisme démocratique, au contraire, exige de l'État qu'il se penche démocratiquement sur cette question en l'affirmant et en la soumettant à discussion, un peu à la manière des articles de Wikipédia aujourd'hui. Dans une telle perspective, il s'agirait donc de conserver une classification unique, mais de la soumettre perpétuellement à discussion. *Cf* BRIEY, Laurent de, 2009. *Le sens du politique: essai sur l'humanisme démocratique*. Editions Mardaga.

l'opposé, les algorithmes des grands acteurs du domaine de l'information sur le net, parce que conçus scientifiquement, permettraient un accès neutre au savoir et laisseraient tout loisir à ses utilisateurs de classer la connaissance et le réel selon leurs propres schémas mentaux.

Mais si l'on soutient au contraire l'idée que la manière avec laquelle le monde accède à l'information aujourd'hui serait moins le résultat d'une évolution technologique récente que celui d'une évolution philosophiques déjà ancienne, quel peut être l'apport réel de la science des données à l'exploration des connaissances aujourd'hui ?

La Classification Décimale Universelle (CDU) à la recherche d'une métaphore visuelle.

En 2013 s'est tenu à La Haye un séminaire[217] organisé par le consortium UDC, c'est-à-dire l'organisme d'éditeurs qui est en charge de la gestion de la Classification Décimale Universelle, cette classification élaborée par les juristes belges Paul Otlet et Henri La Fontaine et dont le but était à l'origine de reprendre la classification décimale de Dewey ainsi que de la rendre plus exhaustive grâce à la création d'indices plus complexes[218]. Portant sur la classification et la visualisation, ce séminaire était le quatrième d'une série d'événements destinés à faire avancer la recherche en matière de classifications bibliographiques, mais aussi à promouvoir un dialogue entre le domaine de la bibliographie et les autres sciences de l'information requérant une organisation de la connaissance. Dans l'introduction du recueil des communications qui en a été publié, les éditeurs écrivent :

> « Récemment, des avancées remarquables ont été faites dans le champ de la visualisation de la connaissance, notamment en relation avec les systèmes d'organisation du savoir dans les sciences, dans les applications de l'extraction de données et dans les tentatives faites pour améliorer l'utilisation de très grands jeux de données et bases de données.
> Le séminaire de 2013 aborde l'enjeu de la visualisation, qui est au cœur du problème de la découverte de l'information et, par conséquent, est un enjeu qui concerne toutes les classifications bibliographiques. L'exploitation médiocre de la classification dans la recherche de l'information a longtemps été attribué au manque de solutions d'interface qui rendrait la complexité de la classification de la connaissance plus facile à présenter et à utiliser pour la navigation dans ces connaissances »[219].

La visualisation des connaissances et de leur organisation, rendue possible notamment par les nouvelles possibilités techniques d'extraction des données, est donc considéré comme un enjeu central pour faciliter la navigation dans l'information et, à l'échelle d'une bibliothèque, dans les collections. Contrairement à la fragmentation de

[217] UDC Seminar 2013, [sans date]. [en ligne]. [Consulté le 16 mai 2014]. Disponible à l'adresse : http://seminar.udcc.org/2013/programme.php

[218] Classification décimale universelle, 2014. *Wikipédia* [en ligne]. [Consulté le 27 août 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Classification_d%C3%A9cimale_universelle&oldid=105773565.

[219] INTERNATIONAL UDC SEMINAR, SLAVIĆ, Aida et UDC CONSORTIUM (THE HAGUE) (éd.), 2013. *Classification & visualization: interfaces to knowledge : proceedings of the International UDC Seminar 24-25 October 2013, The Hague, the Netherlands ; organized by UDC Consortium, The Hague*. Würzburg : Ergon. p. X. « Recently, notable advances have been made in the field of knowledge visualization, especially in relation to knowledge ordering systems in the sciences, in data mining applications and in an attempt to improve the use of large datasets and large databases. The 2013 Seminar addresses the issue of visualization, which is at the heart of the information discovery problem and, by extension, is an issues of concern for all bibliographic classifications. The poor exploitation of classification in information retrieval has been long attributed to the lack of appropriate interface solutions that would make the complexity of knowledge classification easier to present and use in knowlegde browsing ».

l'information, la multi-dimensionnalité des classifications et l'individualisation de l'exploration des connaissances, la visualisation pourrait-être la réelle innovation apportée par le numérique et l'ère du Big Data, à la condition que l'on admette que les métaphores de l'arbre et du labyrinthe ont existé depuis Aristote, mais que la visualisation des connaissances, produite en temps réelle à partir des données bibliographiques et sur le modèle de ces métaphores, est quant à elle nouvelle.

Au vu des nombreuses communications qui ont été faites dans ce séminaire de la CDU, il nous semble que c'est là l'occasion parfaite pour présenter quelques exemples de visualisations de l'organisation des connaissances extraites de ce recueil. Mais auparavant, nous aimerions nous attarder sur les raisons pour lesquelles la visualisation nous paraît être fondamentale pour la navigation dans les collections.

La nécessité d'une métaphore

Dans un des articles introduisant le séminaire de la CDU portant sur la visualisation, on peut lire :

> « Nous avons considéré la division des connaissances en sujets, disciplines ou champs comme une pratique utile déjà bien avant Aristote. Ces divisions sont souvent organisées en métaphores qui, en retour, influencent notre compréhension de la connaissance elle-même. Structurées ou diffuses, se chevauchant ou se séparant, enracinées ou ouvertes, en fractales ou en divisions, ces métaphores nous renseignent sur la manière dont nous pensons la pensée, et elles se prêtent elles-mêmes aux représentations visuelles qui construisent et renforcent nos notions de l'ordre des connaissances »[220].

Scott B. Weingart insiste sur les vertus cognitives des métaphores qui accompagnent depuis toujours l'organisation des connaissances produites sur l'univers. Aristote considérait déjà la métaphore comme une figure de rhétorique ayant à la fois des vertus esthétiques et cognitives : la métaphore doit permettre de rapprocher des objets qui n'ont apparemment rien à voir entre eux afin d'apercevoir des ressemblances ou des affinités entre deux concepts. Umberto Eco prend l'exemple de pirates méditerranéens que l'on qualifierait de pourvoyeurs ou de fournisseurs : le rapprochement nous incite en effet à considérer les pirates non plus sous l'angle moral, mais sous un angle économique que l'on n'aurait pas envisagé auparavant. « Quand Aristote, écrit-il, disait que l'invention d'une belle métaphore "met sous les yeux" pour la première fois un rapport inédit entre deux choses, il voulait dire que la métaphore impose une réorganisation de notre savoir et de nos opinions »[221].

En réalité, l'arbre de Porphyre fait évoluer le statut de la métaphore de représentation mentale et cognitive à celui de représentation visuelle, dont Johanna Drucker a démontré l'utilité dans la diffusion et l'avancement des sciences[222]: en ce qui concerne par exemple les sciences de la terre, pour lesquelles le dessin parfait

[220] WEINGART, Scott B. « From trees to webs : uprooting knowledge through visualization » dans NTERNATIONAL UDC SEMINAR, SLAVIĆ. 2013. p. 43. « Still, we have found the division of knowledge into subjects, disciplines or fields a useful practice since before Artistotle. These divisions are often organized into metaphors, which, in turn, influence our understanding of knowledge itself. Structured or diffuse ; overlapping or separate ; rooted or free, fractals or divisions ; these metaphors inform how we think about thinking, and they lend themselves to visual representations which construct aznd reinforce our notions of the order of knowledge ».

[221] ECO, 2010, p. 88.

[222] « Les images visuelles servent les sciences en usant de propriétés graphiques spécifiques. Les images incarnent l'information à travers trois modes différents, chacun d'entre eux ayant une relation structurelle différente avec leur référent. Elles peuvent fonctionner 1) en offrant une analogie visuelle ou une ressemblance morphologique, 2) en fournissant une image visuelle d'un phénomène invisible, ou 3) en fournissant des conventions visuelles pour structurer des opérations ou des procédures ». DRUCKER, 2010. p. 4.

d'un objet devait permettre de créer une analogie entre un phénomène et sa représentation, elle explique que « la longue liste de distorsions, de dessins de spécimens pour lesquels aucune classification conceptuelle n'était encore établie plaide fortement en faveur des effets des images mentales et de leur influence sur la perception »[223].

La représentation d'un phénomène, notamment grâce à la visualisation des données qu'il peut produire, joue un rôle important dans sa compréhension. Mais où se situe donc la Classification Décimale Universelle par rapport à cela ? Remarquons en premier lieu que Paul Otlet avait conçu sa classification de manière à ce qu'elle présente davantage de souplesse qu'un arbre hiérarchique traditionnel. Il avait en effet fait en sorte qu'elle puisse combiner plusieurs facettes d'un même objet, ce qui rendait sa classification davantage multidimensionnelle que celle de Dewey. Mais au-delà de cela, Otlet a cherché à présenter visuellement cette multi-dimensionnalité et beaucoup de ses illustrations « étaient caractérisées par des représentations non-hiérarchiques de la classification, ressemblant à des réseaux et prévoyant des parcours indirects sans passer par des troncs ou des hiérarchies particulières »[224].

Dans ce contexte, tout l'enjeu du séminaire de la CDU tenu en 2013 était de parvenir à représenter graphiquement la classification et, à l'instar de Paul Otlet son concepteur, d'aller au-delà de la traditionnelle figure arborescente qui caractérisait les classifications précédentes pour pouvoir rendre au mieux la multi-dimensionnalité qui caractérise la CDU et, plus généralement, l'organisation labyrinthique des connaissances.

De l'arbre... à la galaxie.

Au sein du recueil des communications faites à l'occasion du séminaire de la CDU de 2013, nous avons choisi les exemples qui nous paraissaient à la fois les plus liés au monde des bibliothèques et les plus emblématiques des quelques techniques de visualisation dont nous allons exposer ici les caractéristiques.

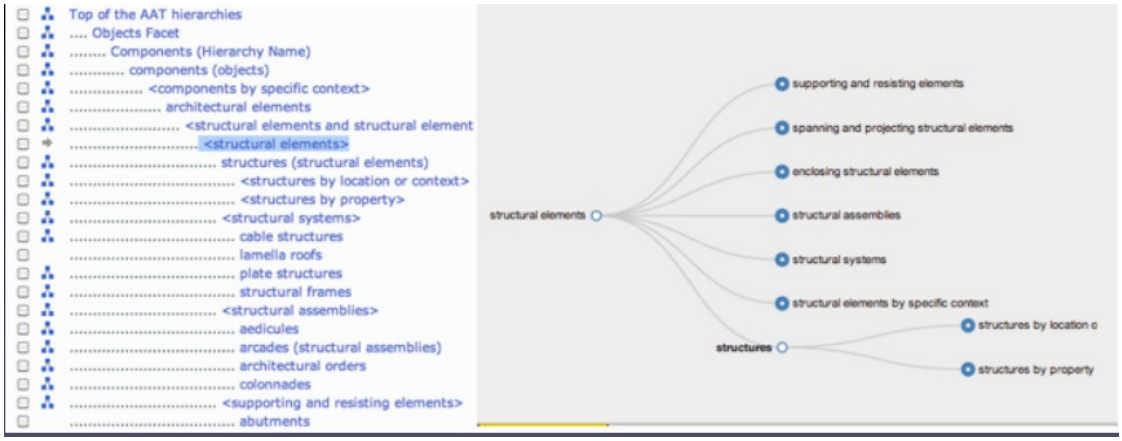


Figure 8 : un exemple de structures hiérarchiques présentées sous forme textuelle et visuelle.

Pour commencer, nous pouvons considérer, au niveau le plus élémentaire de la visualisation des connaissances, la métaphore de l'arbre (figure ci-dessus[225]). Il s'agit simplement de visualiser à la fois des relations entre des concepts mais aussi les

[223] *Ibid*. « The long invetory of distortions, drawings of specimens for which no conceptual classifications is yet established argues strongly for the effects of mental images and their influence on perception ». p. 5.

[224] WEINGART, 2013. « (…) Many of Otlet's illustrations featured non-hierarchical network-like representations of classification, with circuitous paths and no discernible trunk or preferred hierarchy (…) ». p. 50.

hiérarchies qui existent entre eux. La métaphore de l'arbre est bien souvent utilisée par volonté à la fois de simplification et de précision : du fait de ses contraintes hiérarchiques, il n'est pas possible de développer un trop grand nombre de branches à partir d'un terme, ce qui est aussi un avantage car cela permet d'éviter le désordre inhérent à une visualisation en réseau, cette dernière n'ayant ni début ni fin, ni extérieur ni intérieur. C'est une visualisation qui se révèle également peut-être plus facile d'approche, du fait de la familiarité naturelle et universelle que l'on peut entretenir vis-à-vis de la métaphore arborescente. La Classification Universelle Décimale, à l'instar de toutes les classifications traditionnelles, se prête assez bien à une visualisation arborescente, étant donné qu'elle est elle-même conçue sur le modèle de l'arbre : « par exemple, la maladie céliaque en 616.341-008.6, est subordonnée à 616.34, qui est subordonné à 614.3, etc. »[226]. L'arbre apparaît donc comme une représentation naturelle pour la classification, et, par là, la navigation. D'ailleurs, les sites de commerce en ligne l'ont bien compris, puisqu'ils proposent bien souvent une interface à facettes pour naviguer dans leur catalogue de produits (figure ci-dessous[227]), comme l'explique le consortium de bibliothèques universitaires de l'Illinois dans sa page consacrée aux questions fréquemment posées :

> « Les facettes permettent de diviser un ensemble de documents (comme une liste de résultats provenant d'un moteur de recherche) en des sous-ensembles plus petits, à partir d'un élément commun que partagent ces documents. La recherche facettée permet de fournir un moyen à l'utilisateur de restreindre rapidement un ensemble large de documents vaguement liés entre eux en des sous-catégories plus petites. Des exemples populaire de l'utilisation de facettes sont Amazon, eBay et beaucoup d'autres sites d'achats en ligne, de même que quelques catalogues de bibliothèques ou de bases de données d'articles. Par exemple, une recherche pour le terme « chaussures » sur un magasin en ligne vous permet d'avancer en restreignant à chaussures d'homme ou chaussures de femmes, puis de restreindre encore par couleur, prix, etc. »[228].

En clair, les interfaces proposant une recherche à facettes permettent d'affiner une idée comme on pellerait un oignon, et d'avancer dans l'information en partant du général pour arriver au particulier[229]. Il s'agit donc bien là d'une progression hiérarchique et arborescente, le principe de l'arbre étant de cacher les propriétés générales d'un objet derrière des propriétés particulières : la désignation d'une chose comme un chien sous-entend nécessairement qu'elle est aussi un mammifère et un animal. De même, les catalogues de bibliothèques qui proposent une navigation dans leur collection par facettes proposent en réalité ce qui peut être considéré comme la visualisation d'un arbre : l'affichage Primo[230] des

[225] xlin_udcseminar2013.pdf, [sans date]. [en ligne]. [Consulté le 8 septembre 2014]. Disponible à l'adresse : http://www.udcds.com/seminar/2013/media/slides/xlin_udcseminar2013.pdf

[226] RAZPOTNIK, Špela, ŠAUPERL, Alenka. « Enhancing browsing experience through visual presentation of subject terms », dans INTENRATIONAL UDC SEMINAR, SLAVIC, 2013. « e.g. coeliac disease 616.341-008.6, est subordonnée à 616.34, qui est subordonné à 614.3, etc ». p. 212.

[227] *Cf* annexe p. 109, figure 29.

[228] VuFind FAQ: Frequently Asked Questions, [sans date]. [en ligne]. [Consulté le 29 août 2014]. Disponible à l'adresse : http://www.library.illinois.edu/learn/find/vufind/vufind_faq.html. « Facets divide a single set of items (like results from a search engine) into smaller sub-sets based on something those items share in common. Faceted searching provides a way for a user to quickly narrow down a very broad set of loosely related items into smaller sub-sets. Popular examples of the use of facets can be found on Amazon, eBay, and many online shopping sites, as well as some library catalogues and article databases. For example, a search for « shoes » at an online store allows you to narrow further by men's shoes or women's shoes, and then provides further refinements to narrow by color, price, tec. » (Consortium of Academic Research Libraries in Illinois, 2011) ».

[229] LA BARRE, Kathryn. « Sempre avanti ? Some reflections on faceted interfaces », dans INTERNATIONAL UDC SEMINAR, SLAVIC, 2013. p. 94.

[230] *Cf* annexe p. 109, figure 30.

bibliothèques de l'Université de l'Illinois (UIUC) suggère dans son menu un certain nombre de catégories (format, localisation, sujet, auteur, collection, date de publication, etc.), elles-mêmes divisées en sous-catégories (on trouvera dans la catégorie « format » les sous-catégories « articles, périodiques, livres, articles de journaux, etc.). Si l'on prend donc la visualisation dans son sens large d'affichage de l'information, les espaces séparant les catégories et la « mise en gras » des titres de catégories constituent eux-même une visualisation de l'organisation de l'information, de même que l'organisation d'un livre en parties et chapitres.

Figure 9 : Amazon, exemple par excellence d'interface à facettes.

Mais ce type de visualisation, s'il est fort utile, comme l'illustre le succès des sites de ventes en ligne, possède toutefois ses limites lorsqu'il s'agit de visualiser les réseaux complexes de l'information. Les expérimentations conduites à l'Université de l'Illinois ont montré qu'il n'était pas possible d'afficher toutes les sous-catégories attenantes à une recherche, et qu'il était dès lors nécessaire de supprimer des éléments pourtant importants. Par ailleurs, comme le soulignent Xia Lin et Jae-Wook Ahn, « ces structures de connaissances étaient élaborées la plupart du temps par des hommes experts dans chaque domaine et existaient sous la forme de vocabulaires contrôlés et d'ontologies »[231]. On ne voit pas très bien l'utilité, dans ce contexte, d'une visualisation qui se contente de reproduire une représentation déjà déterminée au préalable : elle ne permet pas réellement de découvrir de nouvelles relations entre plusieurs concepts. Kathryn La Barre, qui avait développé l'exemple de la navigation par facettes, en appelle elle-même à découvrir d'autres façons de visualiser l'information[232].

Les nouvelles techniques de traitement des données permettent d'innover dans l'élaboration de visualisations performantes du savoir et de son organisation. « Souvent, écrivent Lin et Ahn, il n'existe pas de structures de connaissances explicite et déjà prête à être visualisée. Dès lors, il faut faire l'effort d'extraire la structure de connaissances de

[231] LIN, Xia, AHN, Jae-WOOK. « Challenges of knowledge structure visualization », dans INTENRATIONAL UDC SEMINAR, SLAVIC, 2013. p. 79.

[232] LA BARRE, 2013. p. 100-101.

données non-structurées en utilisant des techniques variées de fouille de texte[233] avant de pouvoir visualiser les structures ». Les auteurs décrivent ensuite un processus de regroupement (clustering)[234] :

> « Une des méthodes les plus populaires utilisées pour extraire des structures de connaissances est l'algorithme cartographique auto-organisant (…) développé par Teuvo Kohonen (…). L'algorithme utilise un réseau neuronal artificiel qui peut être appris, à partir des vecteurs caractérisant l'ensemble des données textuelles selon les positions de l'extraction de concepts. La carte apprise inclut un nombre de « cellules » qui représente les concepts les plus représentatifs. Les concepts qui leurs sont liés sont calculés à partir du processus d'apprentissage et sont placés dans les cellules avoisinantes »[235].

Les algorithmes de regroupement et de calcul de distances permettent de faire émerger les relations entretenues par plusieurs concepts entre eux et, de là, font apparaître une organisation des connaissances. Les visualisations qui sont produites à partir de ces opérations font apparaître de nouvelles hiérarchies entre les idées et permettraient ainsi de découvrir des associations inconnues jusqu'alors. Ces visualisations sont donc bien souvent en forme de réseaux sémantiques (de « labyrinthes en rhizomes », si l'on voulait employer l'expression d'Umberto Eco) mais ont l'inconvénient d'être assez désordonnées[236]. Il est donc nécessaire de les simplifier au maximum, ce qu'ont proposé Lin et Ahn avec l'Expansion Visuelle de Requête (EVR) dont le principe est de restreindre le réseau d'un concept à ses cinq relations les plus importantes, le chiffre cinq étant choisi arbitrairement[237].

Contrairement à la recherche à facettes développée plus haut, ces visualisations de structures de connaissances élaborées à partir de techniques d'extraction de données, telles que développées dans les exemples que nous venons de citer, ne permettent pas véritablement de naviguer dans les collections d'une bibliothèque, à l'exception peut-être de l'Expansion Visuelle de Requête qui permet à terme de construire une requête selon des opérateurs booléens[238]. Elles permettent simplement de visualiser l'organisation des connaissances et ne sont donc en définitive qu'un appui pour des chercheurs qui voudraient se représenter leur domaine de recherche.

Dans ce contexte, l'exemple de l'utilisation du logiciel d'exploration Tag Galaxy dans un catalogue de bibliothèque (figure ci-dessous[239]), développé par Razpotnik et Šauperl[240], apporte un élément nouveau : de même que l'Expansion Visuelle de Requête, Tag Galaxy donne la capacité de construire visuellement des requêtes complexes en ajoutant un concept à un autre pour restreindre la recherche.

---

[233] *Cf* note n°190.

[234] « Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics ». Cluster analysis, 2014. *Wikipedia, the free encyclopedia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://en.wikipedia.org/w/index.php?title=Cluster_analysis&oldid=63575264. Page Version ID: 635752641

[235] LIN, AHN, 2013. « One of the popular methods used to extract knowledge structures is the self-organizing, mapping alhorithm (SOM) developed by Teuvo Kohonen (Kohonen, 1990). The algorithm makes use of an artificial neural network that can be trained from the feature vectors of the text data set at the positions of the concept extraction (Kaski et al., 1998). The trained map includes a number of "cells" that represent most representative concepts. Related concepts are calculated from the training process and they are placed in the neighbouring cells. » p. 80.

[236] *Cf* annexe p. 110, figure 31.

[237] LIN, AHN, 2013. p. 83. *cf* annexe p. 108, figure 28.

[238] *Ibid*. « This way the user picks up terms to build the query "semantic AND verbal learning AND cognition" to search in *PubMed* ». p. 83.

[239] Tag Galaxy, [sans date]. [en ligne]. [Consulté le 9 septembre 2014]. Disponible à l'adresse : http://taggalaxy.de/

[240] RAZPOTNIK, ŠAUPERL, 2013, p. 216-219.

Cette fois-ci, cependant, nous pouvons descendre jusqu'au niveau du document dont le visuel est lui-même intégré à la représentation[241], contrairement à l'EVR pour lequel la visualisation s'arrête simplement à la construction de la requête. Tag Galaxy est en effet un explorateur visuel utilisé par Flickr afin de permettre à ses utilisateurs de naviguer plus facilement dans l'amas stellaires de photographies quotidiennement postées, taguées et commentées sur le site. Il suffit de rentrer dans la barre de recherche un sujet, comme par exemple « BnF » pour que l'explorateur propose un soleil central représentant ce tag ainsi que plusieurs planètes gravitant autour de ce soleil et représentant les concepts affins. Par un clic sur l'une des planètes, le concept qui lui est attaché s'ajoute à la requête initiale. Un autre clic sur le soleil permet de voir s'afficher en mosaïque sur l'étoile l'ensemble des photographies recherchées, restreint par la liste des concepts qui auront été précédemment sélectionnés parmi les planètes de la galaxie[242]. Razpotnik et Šauperl ont démontré que cet outil était tout à fait adaptable à l'univers d'une bibliothèque[243].

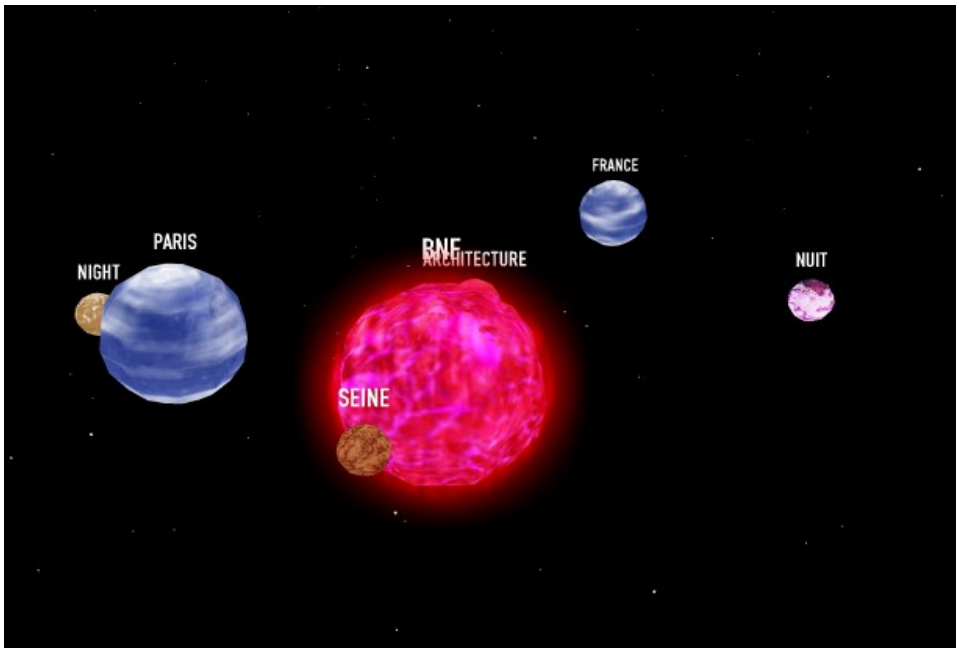


Figure 10 : galaxie se formant autour du tag BnF sur Flickr.

Inutile de dire qu'il s'agit là, à la fois, d'une manière de visualiser l'organisation des connaissances, mais aussi, de naviguer jusqu'à un document recherché. Par ailleurs, les galaxies sont construites par des algorithmes qui donnent mathématiquement une structure aux masses de données qui lui sont soumises : il est donc possible de découvrir, en dehors des hiérarchies déjà connues, des affinités jusque là non envisagées entre plusieurs concepts. Dès lors, Tag Galaxy apporte ceci de nouveau qu'il permet d'opérer une synthèse entre la navigation par des interfaces à facettes, arborescentes, et des opérations effectuées mécaniquement sur des données afin d'en faire émerger des structures invisibles auparavant. Par ailleurs, c'est une manière ludique et agréable de rechercher l'information, même s'il est vrai que le design de l'outil, déjà vieux de quelques années, pourrait être amélioré.

Les technologies affiliées aux données permettent d'aller plus loin encore que la simple navigation et d'explorer virtuellement les collections.

[241] *Cf* annexe p. 110, figure 32.

[242] Démonstration en ligne : *Tag Galaxy - Create Your Own Flickr Photo Universe*, 2011. [en ligne]. [Consulté le 29 août 2014]. Disponible à l'adresse : http://www.youtube.com/watch?v=uDMYByYOCa4&feature=youtube_gdata_player

[243] RAZPOTNIK, ŠAUPERL, 2013.

Rendre visible la bibliothèque sur Internet.

Que peut-être une bibliothèque à l'heure d'Internet et du numérique ? C'est sans doute une banalité de le dire, mais la question de l'identité de la bibliothèque est au cœur de nos[244] préoccupations contemporaines, tant cette dernière est interrogée par les nouveaux usages introduits à l'ère du numérique. Nous tenterons, dans ce dernier chapitre, d'apporter quelques éléments de réponses à cette question, en nous en posant une autre : en quoi la visualisation des données – et avec elles, la visualisation des collections et des publics – pourrait-elle permettre de donner une visibilité nouvelle à la bibliothèque dans le contexte numérique actuel ? Car la question de la représentation est bien, à ce qu'il nous semble, intrinsèquement liée à celle de l'identité.

Nous avons jugé nécessaire de rappeler, en premier lieu, les caractéristiques du nouvel environnement numérique des bibliothèques, ainsi que les problèmes qu'il pose. Nous aimerions démontrer, dans un second temps, la nécessité d'un geste visuel fort pour rendre visible la bibliothèque dont nous aimerions dessiner les principaux traits dans un troisième moment de cette réflexion.

Les bibliothèques dans l'économie de l'attention.

« Le Web a pour effet immédiat de créer une économie de l'abondance d'information (…) », écrit Emmanuelle Bermès :

> « la bibliothèque n'est plus un passage obligé pour accéder aux documents. Toute démarche orientée utilisateur dans l'environnement du Web doit donc prendre en compte comme paramètre premier le besoin de visibilité. La logique du portail est insuffisante sur le Web : c'est dans les moteurs généralistes eux-mêmes, tels que Google, Yahoo !, Bing, etc. qu'il faut gagner en visibilité si l'on veut capter l'attention des internautes.[245] »

« Capter l'attention des internautes », c'est bien là une phrase caractéristique de cette économie dont Emmanuel Kessous a décrit en détail les règles dans son ouvrage intitulé *L'attention au monde : sociologie des données personnelle à l'ère numérique*[246]. L'attention, cette « faculté de l'esprit de se consacrer exclusivement à un objet », y est décrite comme une ressource d'autant plus limitée que l'information est devenue, à l'époque d'Internet, surabondante. Si au XX^e^ siècle, la psychologie a mis en évidence les limites cognitives de l'attention humaine, le XXI^e^ siècle ajoute à cette rareté un accès décuplé à l'information :

> « Si le cyberespace se développe pour englober les interactions entre les milliards de personnes aujourd'hui sur la planète, ces types d'interaction seront totalement différents de ce qui prévalait durant ces derniers siècles ou même avant (…). Lanham rejoint sur ce point Goldhaber et parle à propos d'Internet d'économie pure de l'attention. "Il y a un segment de notre vie actuelle qui constitue une économie de l'attention à l'état pur. Que nous l'appelions cyberespace, virtualité, communication médiée par ordinateur, ou tout simplement le Net, là-bas l'attention est tout. Bien sûr, il y a une foule de signes de retour

[244] L'emploi du « nous » désigne ici à la fois les professionnels de la documentation et le public (ou non) des bibliothèques.

[245] BERMÈS, Emmanuelle, ISAAC, Antoine et POUPEAU, Gautier, 2013. *Le Web sémantique en bibliothèque*. Éditions du Cercle de La Librairie. p. 23.

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

à la "vraie vie", mais ils ne sont que des moyens de sortir d'une économie de l'attention pure" »[247].

Le contexte numérique actuel est donc décrit comme « une économie de l'attention pure » selon les mots de Richard Lanham cités par Kessous. On pourrait objecter à cela que le livre d'Emmanuel Kessous est une sociologie : il se contente de décrire et d'expliquer un ensemble de conception dont il fait une « nouvelle cité »[248], venant s'ajouter aux autres et obéissant à des « principes supérieurs communs »[249]. En ce sens, il ne décrirait pas tant une nouvelle réalité qu'une nouvelle pensée économique. Néanmoins, c'est bien dans ce cadre de l'économie de l'attention qu'Emmanuelle Bermès décrit, très justement à ce qu'il nous semble, les problèmes actuels des bibliothèques. Elle décrit en effet le monde de l'information en utilisant la métaphore du milieu urbain : à l'instar de la bibliothèque physique, implantée dans son environnement urbain, la bibliothèque virtuelle implantée dans l'environnement d'Internet doit se signaler, se rendre visible.

> « Sur le réseau, que l'on peut percevoir comme un vaste espace d'information dans lequel les internautes naviguent en suivant un cheminement qui correspond à leur propre pratique, le site Web de la bibliothèque joue le même rôle que le bâtiment dans la ville. Il doit bien sûr être fonctionnel et immédiatement identifiable. Toutefois, cela n'est pas suffisant car il ne se trouve pas naturellement sur le chemin de l'internaute dans sa navigation : celui-ci va fréquenter son moteur de recherche préféré, la page d'accueil de son fournisseur d'accès, des sites comme Facebook, Wikipédia, Twitter... et c'est à partir de là que se construit sa navigation. Si la bibliothèque ne parvient pas à se rendre visible au sein de ce cheminement naturel, alors il y a toutes les chances que l'internaute passe à côté et utilise d'autres outils pour atteindre ses objectifs, que ceux-ci soient de loisirs, d'apprentissage ou de vie pratique »[250].

À l'instar des visualisations propres à la navigation que nous avons décrit plus haut, c'est de nouveau sous la métaphore de la spatialisation que se décrit l'accès au savoir et à l'information dans l'environnement numérique. Le problème de l'attention à la bibliothèque, de sa visibilité, se pose donc en terme de cheminement urbain. Or, nous avons souligné plus haut l'importance de la métaphore quant à la compréhension de l'organisation de l'information : c'est donc avec la métaphore urbaine que nous nous proposons maintenant de montrer l'intérêt de la visualisation pour la visibilité de la bibliothèque en ligne.

De la monumentalité au geste visuel.

Deux solutions s'offrent au problème de la visibilité de la bibliothèque dans la ville. La première concerne sa situation dans le tissu urbain et la seconde, sa monumentalité, est celle qui la rendrait repérable dans la ville. Emmanuelle Bermès semble pencher pour la première solution :

> « Dans la ville, pour être fréquentée, la bibliothèque a besoin d'être incarnée par un bâtiment visible, immédiatement identifiable pour la fonction qu'il remplit. La bibliothèque met également en place des moyens qui permettent aux lecteurs potentiels de la trouver : des panneaux indicateurs, ou tout simplement un symbole qui positionne son emplacement sur une carte.

[247] *Ibid*. p. 165.
[248] *Ibid*. p. 163.
[249] *Ibid*. p. 155.
[250] BERMÈS, ISAAC, POUPEAU, 2013. p. 24.

> Cependant, cette démarche purement géographique si elle est indispensable, ne peut suffire comme unique moteur pour inciter les gens à venir à la bibliothèque. Les différentes stratégies mises en place par les bibliothèques pour gagner une audience plus large dans la cité, du bibliobus à la bibliothèque hors les murs, en passant par toutes les animations qu'elles organisent (expositions, lectures, accueil des scolaires...) et les différents moyens qu'elles peuvent utiliser pour faire connaître ces activités hors de la bibliothèque (à la mairie, dans d'autres établissements culturels...) participent toutes d'une démarche visant à guider l'utilisateur à la bibliothèque avec pour argument le contenu »[251].

Ainsi, de même que la bibliothèque physique doit sortir de ses murs pour aller à la rencontre de ses usagers, le catalogue en ligne doit pouvoir s'ouvrir aux formats du Web Sémantique pour que les informations qu'il contient puissent apparaître dans les résultats des moteurs de recherche. La présence des bibliothèques sur les réseaux sociaux, comme Facebook et Twitter, participe également de cette idée de rendre visible la bibliothèque par sa présence dans le tissu virtuel. Cependant, dans son mémoire consacré au choix de l'implantation de la bibliothèque dans la ville[252], Grégor Blot-Julienne indique que le positionnement urbain de la bibliothèque ne fait pas tout. Pour lui, « la monumentalité demeure comme un signe » :

> « Outre qu'elle signale la bibliothèque plus efficacement que tous les panneaux, elle seule permet d'intégrer la volonté politique autant que la finalité même de la bibliothèque (…). Et comme la bibliothèque est difficile à définir comme objet d'architecture, elle constitue la monumentalité comme une nécessaire distinction par laquelle elle définit sa place autant que son rôle »[253].

En distinguant la bibliothèque du reste du bâti, le geste architectural contribue à conférer à cette dernière une identité en tant que bâtiment qui, sans cela, resterait difficile à définir. C'est bien là une confirmation du lien intrinsèque qui unit le problème de l'identification du lieu bibliothèque et le problème de sa visibilité. Mais pour poursuivre l'analogie avec la visibilité de la bibliothèque dans l'espace global d'information du web, quel élément numérique pourrait correspondre à la portée du geste architectural, si ce n'est la visualisation de données ? Il nous semble en effet significatif que Johanna Drucker, dans sa tentative pour définir ce qu'elle appelle *graphesis*, se réfère aux artistes du Bahaus et plus particulièrement à Kandinsky dont la conviction était que le design convenait à tout média et à toute discipline, y compris ce qui deviendrait plus tard l'infographie. De fait, Johanna Drucker fait du design un élément central de la visualisation et, par extension, de tout média numérique[254]. Or, ce caractère central du design dans la visualisation nous paraît tout à fait fondamental lorsque l'on considère quelle place les économistes de l'attention lui accordent eux-mêmes dans leur « cité ». « Le design d'un produit, écrit Richard Lanham, nous invite à lui prêter attention de manière particulière, à porter un certain type d'attention sur lui.

[251] *Ibid*.

[252] BLOT-JULIENNE, Grégor, 2012. *Du choix de l'implantation aux stratégies de localisation: bibliothèques dans la ville*. Bibliothèque numérique de l'Enssib. Consulté le 30 août 2014. Disponible à l'adresse Web : http://www.enssib.fr/bibliotheque-numerique/documents/56709-du-choix-de-l-implantation-aux-strategies-de-localisation-bibliotheques-dans-la-ville.pdf

[253] *Ibid*. p. 21.

[254] DRUCKER, 2010. « Digital technology depends on visual presentation for much of its effectiveness. The ubiquitous graphical user interface and design of icons for navigation, daily activities and functions are, familiar graphic structures. Higher-level functions using visualization are commonplace for analysis of statistical data. Many creative, original works in all areas of design for industry, art, entertainment, engineering, and technological activity at micro and macro levels are graphically enabled through design ». p. 1.

Le design ne nous dit rien sur les choses elles-mêmes, mais sur ce que nous pensons des choses. Il est l'interface où les chose que nous extrayons de la croûte terrestre rencontrent une réalité pleinement humaine de sentiments, d'attitudes, et d'ambitions »[255].

Et si, à l'instar d'Emmanuelle Bermès, nous parlions de catalogue en ligne de bibliothèque et que nous lui ajoutions une dimension visuelle par l'intermédiaire des données qu'il contient, nous ajouterions là un autre élément fondamental de l'économie de l'attention, à savoir la faculté de filtrer l'information, et éventuellement d'effectuer des recommandations, – ce qui d'ailleurs a de plus en plus tendance à se développer dans les catalogues « nouvelle génération » –. Nous aurions là un dispositif numérique tout à fait efficace dans l'environnement virtuel[256]. Reste à savoir sous quelle forme nous pourrions décliner ce dispositif : c'est ce que nous allons tenter de définir maintenant.

Un data game stellaire ?

Un data game, ou jeu avec des données, est un jeu vidéo dont l'environnement et le scénario sont uniquement fondés sur des données réelles. Le principe du data game est simplement d'être un jeu vidéo, une simulation, dont le contenu est apporté par des données se référant à des objets réels. Ainsi, dans le cas d'un jeu sérieux (serious game) dont le but serait de simuler des explosions de bombes nucléaires et d'en mesurer les conséquences, celui-ci se déclinerait ainsi :

> « Le lieu, d'abord, pourra ainsi être une grande ville contemporaine ou un des endroits de la planète qui a déjà connu une explosion atomique. Et la puissance de la bombe, elle, pourra correspondre à l'arme larguée sur Hiroshima, à la plus grosse ogive française ou encore à la « Tsar bomba », mastodonte de l'armée soviétique. (…) Quant à l'objectif, il est suggéré par le dispositif : recréer toutes les conditions de l'attaque sur Nagasaki et voir combien de victimes elle ferait si elle avait lieu aujourd'hui, déterminer quelle cible serait la plus intéressante pour les États-Unis si nous étions encore en guerre froide, voir jusqu'où s'étendrait la contamination provoquée par une bombe nord-coréenne touchant Séoul... »[257].

Si le principe de ce type de jeu est de se fonder sur des données d'objets réellement existants (la population d'une ville, le nombre de ses voies de communications, etc.), cela signifie que l'on pourrait tout aussi bien choisir les données bibliographiques d'un OPAC et proposer un jeu dont le but serait de les explorer. La vertu principale du data gaming est en effet de proposer une exploration interactive des données et de leur caractéristiques : « Dans un jeu, il y a le plus souvent un conflit, une opposition, un obstacle à comprendre pour mieux le surmonter. (…). À vous de construire une expérience, en proposant des choix qui seront intéressants pour votre public. Ainsi, vous l'amènerez à explorer, à analyser et à comparer des éléments afin de prendre la meilleure décision possible pour atteindre l'objectif fixé »[258]. Explorer, analyser, comparer... Voilà un outil de médiation fort intéressant pour qui veut maintenir pour longtemps l'attention d'un public diversifié sur le catalogue en ligne d'une bibliothèque, ce « reflet de la collection physique »[259], comme l'écrit Emmanuelle Bermès : un data game dont les données seraient celles d'un catalogue permettrait donc d'explorer virtuellement les collections d'une bibliothèque. Mais il est également question ici d'« objectifs » : c'est là

[255] Richard Lanham dans KESSOUS, 2012. p. 170.

[256] *Ibid.* « Les filtres et les moteurs de recherche ou de recommandation (…) constituent un moyen de répondre aux phénomènes de surcharge cognitive mais permettent aussi des ouvertures exploratoires, en jouant sur la curiosité ». p. 173.

[257] Du jeu de données au jeu avec les données | The Pixel Hunt, [sans date]. [en ligne]. [Consulté le 30 juillet 2014]. Disponible à l'adresse : http://florentmaurin.com/?p=471

[258] *Ibid*.

[259] BERMÈS, ISAAC, POUPEAU, 2013.

le propre de la « ludification ». Le jeu permet en effet l'apprentissage, l'appropriation d'un objet complexe par le biais d'une expérience interactive, obéissant à des règles, astreinte à un ou plusieurs objectifs particuliers et produisant des résultats variables en fonction de l'action du joueur : il suffirait alors de masquer l'exploration des données du catalogue derrière un scénario qui s'appuierait sur une métaphore. Il reste à savoir quelle pourrait être cette métaphore : Umberto Eco avait souligné la conformité de la métaphore du cosmos avec la représentation d'une organisation du savoir. Cette même métaphore ayant été développée dans l'exemple de Tag Galaxy présenté plus haut, pourrait s'appliquer à un data game élaboré à partir des données d'un catalogue de bibliothèque. Chaque livre ou chaque auteur ou chaque sujet particulier pourrait être représenté par une planète, appartenant à un système qui serait lui-même constitué par l'ensemble des livres(-planètes) que l'ordinateur aurait jugé proches selon les critères que l'on souhaite (selon son sujet, son auteur, son ambiance, sa couleur la plus présente, etc...), ces systèmes étant à leur tour regroupés en galaxies.

Figure 11 : Présentation du knowledge graph *de Google. Ici, les femmes et le prix Nobel.*

L'exploration des collections de la bibliothèque serait alors développée selon le thème de l'exploration spatiale, un thème qui, nous semble-t-il, est relativement universel car le ciel étoilé est un bien commun et familier de tous. Ajoutons qu'il s'agit d'un thème transgénérationnel, à la fois sérieux si l'on pense aux derniers projets en date de la NASA concernant la planète Mars, mais aussi divertissant si l'on se réfère aux blockbusters les plus récents. Citons d'ailleurs le générique des dernières adaptations cinématographiques de *Star Trek*, portées sur le grand écran par le réalisateur Jeffrey Jacob Abrams, qui nous paraît illustrer ce que pourrait être l'exploration spatiale d'un catalogue de bibliothèque [260]: à l'instar de la caméra, le lecteur avancerait de galaxies en galaxies (formées, rappelons-le, par les calculs d'un ordinateur), les noms des acteurs étant remplacés par des titres de livres, ou par des noms d'auteurs. C'est d'ailleurs un modèle que Google propose, à sa manière, en introduction à son *knowledge graph* (figure ci-dessus[261]). Chaque entité est en effet représentée sous la forme d'un nœud, connectée à d'autres entités

[260] *Star Trek: Into Darkness - End Credits: Preview (2013) | SD*, [sans date]. [en ligne]. [Consulté le 30 juillet 2014]. Disponible à l'adresse : http://www.youtube.com/watch?v=_-W6XIWOiMA

qui gravitent autour d'elle, sur un fond étoilé : il ne s'agit cependant pas d'un outil de navigation, mais d'une simple illustration de la structure encyclopédique du knowledge graph. L'OCLC propose une visualisation similaire, non pas cantonnée cette fois-ci à un rôle d'illustration, mais présentée comme une véritable expérience interactive : un clic sur les auteurs présentés sous la forme de nœuds permet de visualiser leurs relations[262]. Cependant, cela ne concerne que des auteurs ou créateurs, et ne permet pas d'aboutir directement aux documents, ni de construire véritablement une requête.

L'avantage du jeu vidéo, dans ce contexte, est de présenter le catalogue de la bibliothèque de manière ludique : ni véritablement catalogue, ni véritablement bibliothèque numérique, il serait pensé comme un outil de médiation dont l'intérêt principal serait sa visibilité et son attraction cognitive. Ajoutons que l'intérêt du jeu vidéo est également qu'il peut être joué à plusieurs : il peut donc être utilisé pour animer une communauté. Cette communauté pourrait être constituée de lecteurs, et dès lors, la bibliothèque remplirait en quelque sorte son rôle politique, à savoir celui de concevoir un projet d'accès commun à la connaissance. Mais puisque par ailleurs il s'agit bien d'un jeu, quel pourrait être l'objectif qui lui serait assigné ? Nous pourrions penser à une compétition portant sur la distance parcourue par un lecteur dans l'univers d'une bibliothèque : le but serait ainsi de conduire le lecteur à étendre ses horizons et à rechercher la distance qui peut exister entre plusieurs livres. Des ponts existeraient d'une galaxie à une autre, là où le lecteur aurait remarqué des similarités peut être inattendues entre plusieurs livres. Le bibliothécaire pourrait lui-même jouer virtuellement son rôle de médiation en orientant, conseillant, recommandant des ouvrages, suggérant des passerelles, etc. Par là, nous répondrions peut-être d'une certaine manière à la nécessité contradictoire de proposer « un langage commun, qui garantisse une autonomie des usagers, tout en répondant à leurs points de vue multiples »[263]. Un exemple dont la portée s'étend au-delà du monde des bibliothèques en est « l'incroyable We are Data[264], modélisation interactive de toutes les données qui nous entourent quand on vit dans une grande ville comme Paris (…) »[265]. Cette application web a été développée par Ubisoft sur le modèle d'un jeu vidéo qu'elle est censée promouvoir à partir des données réelles de trois grandes villes. On voit par là que la production d'un jeu vidéo interactif permettant d'explorer les collections d'une bibliothèque n'est donc clairement pas hors de portée.

Nouveau modèle de bibliothèque ou renouvellement d'un modèle de bibliothèque ?

En interrogeant l'apport réel du numérique et du Big Data à l'accès à la connaissance, nous avons cherché à proposer le raisonnement suivant : en premier lieu, nous postulons que l'évolution d'un accès arborescent et hiérarchique à la connaissance à un accès fragmentaire et multidimensionnel ne serait pas lié à l'apport des nouvelles technologies mais à cette évolution qui a fait que nous sommes passé de l'univers clos de l'antiquité aux perspectives infinies ouvertes par la révolution copernicienne. En somme, la métaphore des feuilles de l'arbre utilisée par Weinberger n'est pas bonne pour désigner l'organisation des connaissances à notre époque. Il faudrait au contraire

[261] *Introducing the Knowledge Graph*, 2012. [en ligne]. [Consulté le 2 septembre 2014]. Disponible à l'adresse : http://www.youtube.com/watch?v=mmQl6VGvX-c&feature=youtube_gdata_player

[262] *Cf* annexe p. 111, document n°5.

[263] TESNIÈRE, Valérie, dans BERTRAND, Anne-Marie, BETTEGA, Emilie, CLÉMENT, Catherine, ERMAKOFF, Thierry, EVANS, Christophe, ION, Christina, PICARD, David-Georges, RAPATEL, Livia, TESNIÈRE, Valérie. *Quel modèle de bibliothèque?*, 2008. Presses de l'ENSSIB. p. 146.

[264] Watch_Dogs WeAreData, [sans date]. *Watch_Dogs WeAreData* [en ligne]. [Consulté le 31 août 2014]. Disponible à l'adresse : http://wearedata.watchdogs.com/. *Cf* annexe p.111, document n°4.

[265] Du jeu de donnée au jeu avec les données. [sans date].

reprendre la métaphore du cosmos employée par Umberto Eco pour désigner l'exploration des connaissances aujourd'hui. C'est la raison pour laquelle nous pensons que le réel apport du Big Data à l'accès à l'information ne réside pas dans les moteurs de recherches ou les systèmes de recommandation, mais bien dans la mise en avant, par la visualisation, de l'aspect métaphorique intrinsèque à toute exploration des connaissances : l'image, et en particulier l'image du labyrinthe ou du cosmos, est particulièrement en mesure de permettre la découverte de connexions nouvelles entre des domaines de connaissance pourtant auparavant éloignés.

En définitive, ce que nous avons cherché à décrire dans le troisième temps de cette étude, est peut-être une réflexion sur les modèles de bibliothèque : modèles d'organisation et de classification des connaissances, modèles de visualisation de ces organisations et, en dernier lieu, modèles de navigation dans les collections, virtuelles, mais aussi physiques.

Envisager l'exploration des collections sur le thème de l'exploration spatiale, ce serait se faire l'écho des remarques de Valérie Tesnière sur la bibliothèque envisagée comme espace de la collection[266] : nous avons souhaité en effet nous demander de quelle manière l'interaction des lecteurs avec les collections, par l'intermédiaire des données et de leurs techniques, pouvait contribuer à définir l'espace de la bibliothèque virtuelle et construire un véritable dialogue entre monde numérique et monde physique. Nous passerions donc de la bibliothèque comme espace de la collection à la collection comme espace de la bibliothèque.

[266] TESNIERE, 2008. « (…) Face à une inéluctable hybridation de la bibliothèque par la fonction de documentation, la notion de collection est-elle bloquante ou bien pense-t-on ceci parce que le rapport collection/bibliothèque reste mal défini ? C'est bien là que l'on perçoit l'ambiguïté profonde du terme "bibliothèque numérique" comme référence à un lieu ou à un contenant, avant que d'être la référence au contenu, à savoir la collection.

Or dans les représentations collectives de la bibliothèque, pas toujours explicitées, il y a, en effet, attaché à la bibliothèque, quelque chose qui résiste, qui concerne la collection en tant qu'outil public. C'est là, me semble-t-il, le sens de l'attachement du public : le nier serait se couper du sens commun ». p. 144-145.

CONCLUSION : DONNÉES ET POLITIQUE

Nous n'avons pu nous empêcher de constater, au cours de la rédaction de cette étude, la proximité étroite qui existe entre les données des bibliothèques et de l'information et ce que l'on pourrait appeler le politique, au sens de vision globale du monde organisant la vie de la cité. L'insistance sur la volonté de neutralité qui se lit dans les discours accompagnant l'accès à l'information à l'ère du Big Data[267], qu'il s'agisse de ses aspects épistémologiques à travers une certaine lecture de la science des données, de la fourniture de contenus à travers la PDA, ou de l'organisation de ces contenus à travers le refus d'une classification générale et collective au profit d'un accès individuel et fragmenté à l'information, nous paraît caractériser ce qui précisément est par nature politique, à savoir l'information et cet objet de diffusion de la connaissance qu'est la bibliothèque. Comment ne pas penser en effet, que cette revendication de la neutralité est d'autant plus forte que l'absence du politique est loin d'être une évidence quand il s'agit de l'information en contexte démocratique ?

Dès lors, si nous voulons nous donner la peine de relire la progression de notre réflexion à la lumière de cet aspect politique essentiel qui lui est attaché, trois axes se dégagent, ou plutôt, trois questions se posent : comment mieux connaître cet objet politique qu'est la bibliothèque et, à travers elle, l'information ? Comment piloter la bibliothèque quand celle-ci est au cœur d'un dialogue permanent entre un trio d'acteurs que sont le bibliothécaire, son élu et ses usagers ? Et enfin, comment fournir un accès pour tous à la connaissance et aux collections de la bibliothèque quand celui-ci est nécessairement le résultat d'une vision propre à un groupe dominant concernant le monde et son organisation ?

« L'histoire des bibliothèques, comme leurs professionnels, souffre de corporatisme », écrivait Martine Poulain :

> « Cette histoire est en effet encore insuffisamment liée à l'histoire culturelle, sociale, politique générale des sociétés et des époques auxquelles elles appartiennent et dont elles sont nécessairement un miroir et un reflet. Quoi de plus nécessairement politique, pourtant, dans toute l'histoire des sociétés que l'histoire des conceptions du livre, de l'écrit et de leur partage »[268] ?

Si l'on veut donc concevoir la bibliothèque et ses (méta)données comme un objet politique, miroir de l'évolution des normes et des valeurs avec lesquelles elles interagit continuellement, il devient nécessaire de penser un outil de connaissance qui confère à ce caractère politique une place centrale. C'est précisément ce que tentent de faire les Humanités Numériques, qui ont voulu élaborer un cadre critique pour la visualisation des données et ont fait de la subjectivité inhérente à cette dernière le point de départ d'une connaissance de la bibliothèque et de son histoire : en témoigne l'excellent essai *The life and death of metadata*[269]. Il n'est pas à exclure que les algorithmes puissent être utilisés de la même manière, et d'ailleurs, ils le sont déjà dans une certaine mesure, puisque d'une part, nous avons vu que des sociologues se sont attachés à décortiquer leurs présupposés, et d'autre part, la visualisation repose largement sur ces algorithmes. Néanmoins, il nous

[267] Google, dans la page de présentation de son équipe, se revendique comme étant une démocratie, classant ses pages par la mécanique objective des liens hypertextes : « 4. La démocratie sur le Web fonctionne » dans 10 principes fondamentaux – Société – Google, [sans date]. [en ligne]. [Consulté le 11 décembre 2014]. Disponible à l'adresse : http://www.google.fr/intl/fr/about/company/philosophy/

[268] POULAIN, Martine, 2002. Retourner à Tocqueville. [en ligne]. 1 janvier 2002. [Consulté le 11 décembre 2014]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2002-05-0066-001/2002/5/fam-apropos/varia

[269] The Life and Death of Data, [sans date]. *op. cit.*

semble que la différence fondamentale entre algorithmes et visualisation de données demeure dans le fait que la subjectivité est latente dans les premiers, tandis que dans la seconde, elle est davantage affirmée : à ce qu'il nous semble, il y a en effet une différence notoire entre se contenter de transposer une vision du monde dans un média et vouloir exprimer, certes médiatiquement, mais également métaphoriquement, cette même vision du monde.

Dans un second temps, si l'on observe la bibliothèque et ses collections du point de vue de leur pilotage, de nouveau, les données posent la question fondamentale du politique, notamment en raison de la particularité des acteurs qu'elles impliquent. Ces acteurs sont décrits par Benoît Tuleu :

> « Dans un contexte nouveau où les missions pédagogiques de la bibliothèque seraient enfin garanties, on aurait tout à gagner à inventer un nouveau triangle élu/bibliothécaire/usager, et à placer en son centre la bibliothèque comme objet politique fondamentale »[270].

Si donc la bibliothèque doit être conçue comme étant au cœur d'une négociation permanente entre un bibliothécaire, son élu et les usagers de son service, il apparaît nécessaire de réfléchir à un outil de pilotage qui permette d'intégrer ce caractère de dialogue continu. À cet égard, la visualisation nous paraît de nouveau être une possibilité intéressante : par son caractère métaphorique, elle rend bien compte du caractère fondamentalement symbolique des variables qui sont choisies pour représenter son activité. Elle est par ailleurs un moyen ludique d'apprendre à faire parler les données, tout en faisant sentir à l'apprenant le caractère construit de ce langage, ne serait-ce que parce que cet apprentissage implique de réfléchir au choix d'une métaphore pour exprimer la bibliothèque. Par ailleurs l'aspect instable des données ou des variables pensées comme des symboles permet également d'envisager que le sens qui est construit à partir des données ne peut être fixe et certain : la confrontation des variables entre elles, telle que recommandée par Jamene Brooks-Kieffer, implique de questionner de manière permanente le sens des données.

De ce fait, la bibliothèque, par le biais de ses données, est au cœur d'un dialogue permanent avec l'élu. Or, par ces qualités communicationnelles, la visualisation est ce qui, par excellence, permet d'enclencher et de renouveler régulièrement les termes de ce dialogue avec les tutelles de la bibliothèque. Mais quid, dans ce contexte, du dialogue avec les usagers de la bibliothèque ? La PDA, il est vrai, a permis de nourrir l'espoir d'une participation intégrale des usagers dans le pilotage de la collection d'un établissement, par le biais notamment de moteurs de recherches et d'acquisitions à la consultation. Cependant, il apparaît nécessaire de penser les limites d'un tel système : si le pilotage d'une bibliothèque doit se faire par le biais d'algorithmes utilisés par des usagers, et si ces algorithmes sont bien des médias, comme nous l'avons écrit dans notre première partie, alors la PDA reviendrait, dans ces circonstances, à « médiatiser » la bibliothèque, à savoir la transformer en un moyen de communication où se refléteraient les opinions dominantes, fondées ou non, des utilisateurs de la bibliothèque. De fait, si la bibliothèque devait être intégralement conçue sur ce modèle, elle serait davantage amenée à exprimer les préférences d'un certain public plutôt qu'à véritablement proposer à des citoyens un projet relatif à l'information : pourrait-on dans ces conditions parler d'un dialogue entre le bibliothécaire et ses usagers ?

Si donc l'on voulait maintenir ce dialogue permanent à l'ère des mégadonnées et leurs technologies, comment envisager les données des bibliothèques en faisant

[270] TULEU, Benoît, 2011. Trop loin, trop proche. [en ligne]. 1 janvier 2011. [Consulté le 11 décembre 2014]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2011-02-0014-002

en sorte que l'aspect politique de la bibliothèque comme lieu de connaissance soit affirmé et soumis à une discussion ? « L'acte fondateur d'une bibliothèque », écrit Benoît Tuleu, « est toujours un geste politique » :

> « (…) En France, historiquement, le chef politique est quelqu'un qui veut laisser sa trace historique dans la pierre des bâtiments et, si possible, dans celle des monuments publics. La bibliothèque joue ce rôle, mais aussi participe au dessin d'un centre-ville, structure un quartier, devient un repère toponyme pour les habitants qui en sont donc tous un peu usagers »[271].

Avec Benoît Tuleu, nous aimerions affirmer que tout édifice d'accès à la connaissance, qu'il soit physique ou virtuel, monumental ou algorithmique, comporte un présupposé politique qu'il transpose dans cet édifice. De ce fait, même à l'ère d'internet et d'une navigation individuelle à travers une information fragmentée, la subjectivité qui autrefois était inhérente à la classification est déplacée au niveau des algorithmes et des adaptations et appropriations qu'ils impliquent chez leurs utilisateurs quant à leur comportement de recherche[272]. En conséquence, de même que le geste architectural, à travers la construction d'une bibliothèque, avait pour rôle d'exprimer et de rendre visible une vision politique sous-jacente à un projet relatif à la connaissance, de même, le geste visuel peut-il permettre d'exprimer métaphoriquement dans l'espace virtuel cette vision politique, tout en lui conférant une visibilité par la captation de l'attention des internautes.

C'est bien là le rôle que peut remplir, nous semble-t-il, un data game dont le contenu serait les métadonnées d'une ou de plusieurs collections. Ce jeu vidéo fonctionnant sur les données d'un catalogue permettrait ainsi de donner un support allégorique à l'accès à la connaissance, tout en refusant de prendre trop au sérieux cette allégorie, en lui reconnaissant son caractère intrinsèque de média. Le principe fondateur d'un jeu vidéo n'est-il pas de simuler ? A ce titre, nous pourrions, par exemple, faire comme si la connaissance s'organisait en un vaste cosmos, comme si le lecteur était un cosmonaute explorant les nouveaux horizons insoupçonnés du savoir, et comme si le bibliothécaire était lui-même un guide dans cet univers fictif : le « comme si » étant la clé conventionnelle devant éveiller et susciter la réaction de ceux à qui il s'adresse.

De la sorte s'élaborerait, avec un nouveau modèle de bibliothèque, un nouveau sens du politique : là où Google, en héritier du libéralisme politique, se contentait de définir une organisation « juste » de la connaissance par le biais de son classement, tout en ne reconnaissant pas nécessairement les valeurs qui constituent malgré lui cette organisation, il s'agirait désormais de proposer une organisation « bonne », c'est-à-dire porteuse d'un sens construit démocratiquement, tout en donnant au citoyen les moyens de reconnaître et de discuter le bien-fondé des principes sous-tendant cette organisation. En cela le mouvement du Big Data apporterait quelque chose de radicalement nouveau.

[271] *Ibid*.

[272] Ainsi Ronald E. Day explique-t-il à propos du Science Citation Index, principe bibliométrique à l'origine même du fonctionnement de l'algorithme de Google. DAY. 2014. « That a small number of authors publish a greater number of works is a sociological fact, not a bibliometric one. It belongs to the logic and distributions (the "grammars") of social power in particular types of sociocultural systems. Feeding this back into the production system in terms of social rewards or in temrs of favored search term leads to exponentially increasing the powers of the sociological systems and does little for the more marginal or unrepresented authors and works that were present (or not) for counting in the first place. ». p. 70.

Bibliographie

ARTICLES ENCYCLOPÉDIQUES

Big data, 2014. *Wikipedia, the free encyclopedia* [en ligne]. [Consulté le 1 novembre 2014]. Disponible à l'adresse : http://en.wikipedia.org/w/index.php?title=Big_data&oldid=631791921. Page Version ID: 631791921

Classification décimale universelle, 2014. *Wikipédia* [en ligne]. [Consulté le 27 août 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Classification_d%C3%A9cimale_universelle&oldid=105773565.

Data set, 2014. *Wikipedia, the free encyclopedia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://en.wikipedia.org/w/index.php?title=Data_set&oldid=625099781. Page Version ID: 625099781

Fouille de textes, 2014. *Wikipédia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Fouille_de_textes&oldid=107660108. Page Version ID: 107660108

Indicateur, 2014. *Wikipédia* [en ligne]. [Consulté le 9 novembre 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Indicateur&oldid=106207898. Page Version ID: 106207898

Ontologie (informatique), 2014. *Wikipédia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Ontologie_(informatique)&oldid=109058774. Page Version ID: 109058774

Représentation graphique de données statistiques, 2014. *Wikipédia* [en ligne]. [Consulté le 12 décembre 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Repr%C3%A9sentation_graphique_de_donn%C3%A9es_statistiques&oldid=108854835. Page Version ID: 108854835

Rhizome (philosophy), 2014. *Wikipedia, the free encyclopedia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://en.wikipedia.org/w/index.php?title=Rhizome_(philosophy)&oldid=637871872. Page Version ID: 637871872

Spécifications fonctionnelles des notices bibliographiques, 2014. *Wikipédia* [en ligne]. [Consulté le 4 août 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=Sp%C3%A9cifications_fonctionnelles_des_notices_bibliographiques&oldid=103576162.

SPSS, 2014. *Wikipédia* [en ligne]. [Consulté le 12 décembre 2014]. Disponible à l'adresse : http://fr.wikipedia.org/w/index.php?title=SPSS&oldid=109086133. Page Version ID: 109086133

Text mining, 2014. *Wikipedia, the free encyclopedia* [en ligne]. [Consulté le 14 décembre 2014]. Disponible à l'adresse : http://en.wikipedia.org/w/index.php?title=Text_mining&oldid=637280039. Version ID: 637280039

MÉMOIRES

BAUDIÈRE, Marie, 2013. *Le bibliothécaire, son élu, son directeur Marie Baudière.* Bibliothèque numérique de l'Enssib. Consulté le 20 août 2014. Disponible à l'adresse Web : http://www.enssib.fr/bibliotheque-numerique/documents/64142-le-bibliothecaire-son-elu-son-directeur.pdf.

BLOT-JULIENNE, Grégor, 2012. *Du choix de l'implantation aux stratégies de localisation: bibliothèques dans la ville*. Bibliothèque numérique de l'Enssib. Consulté le 30 août 2014. Disponible à l'adresse Web : http://www.enssib.fr/bibliotheque-numerique/documents/56709-du-choix-de-l-implantation-aux-strategies-de-localisation-bibliotheques-dans-la-ville.pdf

CARTIER, Aurore, 2012. *Bibliothèque et Open data. Et si on ouvrait les bibliothèques sur l'avenir ?* Consulté le 15 décembre 2014. Disponible à l'adresse Web : http://www.enssib.fr/bibliotheque-numerique/documents/60401-bibliotheque-et-open-data-et-si-on-ouvrait-les-bibliotheques-sur-l-avenir.pdf. p. 61.

GAILLARD, Rémi, 2013. *De l'Open data à l'Open research data quelle(s) politique(s) pour les données de recherche ?* Bibliothèque Numérique de l'Enssib. Consulté le 18 août 2014. Disponible à l'adresse Web : http://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf

TISSERANT, Clément, 2013. *Domaine public et biens communs de la connaisance.* Sous la direction de Cristina Ion. Disponible à l'adresse Web : http://www.enssib.fr/bibliotheque-numerique/documents/64245-domaine-public-et-biens-communs-de-la-connaissance.pdf

MONOGRAPHIES

ALEMBERT, Jean Le Rond d' et CONDORCET, Jean-Antoine-Nicolas de Caritat marquis de, 1821. *Œuvres de d'Alembert*. A. Belin. Volume 1, p. 44.

ALONZO, Valérie, RENARD, Pierre-Yves (dir.). 2012. *Évaluer la bibliothèque*. Bibliothèques (Paris. 1978), 0184-0886

BATTLES, Matthew. 2013. « Data artefacts : tracking knowledge-ordering conflicts through visualization. » dans INTERNATIONAL UDC SEMINAR, Slavić, Aida et UDC CONSORTIUM (THE HAGUE) (éd.), 2013. *Classification & visualization: interfaces to knowledge : proceedings of the International UDC*

Seminar 24-25 October 2013, The Hague, the Netherlands ; organized by UDC Consortium, The Hague. Würzburg : Ergon. ISBN 9783956500077 3956500075.

BERMÈS, Emmanuelle, ISAAC, Antoine et POUPEAU, Gautier, 2013. *Le Web sémantique en bibliothèque*. Éditions du Cercle de La Librairie. ISBN 9782765414179.

BORGES, Jorge Luis. 1944. « Pierre Ménard, auteur du *Quichotte* » dans *Fictions.* Éditions Gallimard. ISBN 9782070366149.

BRIEY, Laurent de, 2009. *Le sens du politique: essai sur l'humanisme démocratique*. Editions Mardaga. ISBN 9782804700102.

CUKIER, Kenneth, MAYER-SCHOENBERGER, Viktor et DHIFALLAH, Hayet, 2014. *Big Data*. Paris : ROBERT LAFFONT. ISBN 9782221140048.

BROOKS-KIEFFER, Jamene. « Yielding to persuasion : Library Data's Hazardous Surfaces » dans ORCUTT, Darby, 2010. *Library Data: Empowering Practice and Persuasion*. ABC-CLIO. ISBN 9781591588269.

CRAMER, Florian, CUBAUD, Pierre, DACOS, Marin, JAMES, Yannick, LANTENOIS, Annick (dir.). 2011 *Lire à l'écran : contribution du design aux pratiques et aux apprentissages des savoirs dans la culture numérique : [actes de la journée d'étude Lectures numériques, Valence, 11 mars 2010]*. Organisée par l'École supérieure d'art et design Grenoble-Valence.

DAY, Ronald E. « "The Data – It is *Me* !" ("Les données – c'est *Moi !*") » dans CRONIN, Blaise et SUGIMOTO, Cassidy R., 2014. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Cambridge, Massachusetts : MIT Press. ISBN 9780262525510.

DELCARMINE, Nadine. « Tableaux de bord en bibliothèque » dans ALONZO et RENARD, 2012.

ECO, Umberto. 1965. *L'œuvre ouverte*. Collection « Points », Éditions du Seuil, Paris.

ECO, Umberto, 2010. *De l'arbre au labyrinthe études historiques sur le signe et l'interprétation*. Paris : Grasset. 978-2-246-74851-9

ELGUINDI, Anne C., MAYER, Bill. « Telling your library's story : how to make the most of your data in a presentation » dans ORCUTT, 2010

EVANS, Christophe (dir). *Mener l'enquête : guide des études de publics en bibliothèque*. 2011. Collection La boîte à outils, 1259-4857

FARGE, Arlette., 1997. *Le Goût de l'archive*. [Paris] : Seuil. ISBN 2020309092 9782020309097.

KELLAM, Lynda M et PETER, Katharin, 2011. *Numeric data services and sources for the general reference librarian*. Oxford : Chandos Publishing. ISBN 1843345803 9781843345800.

INTERNATIONAL UDC SEMINAR, SLAVIĆ, Aida et UDC CONSORTIUM (THE HAGUE) (éd.), 2013. *Classification & visualization: interfaces to knowledge : proceedings of the International UDC Seminar 24-25 October 2013, The Hague, the Netherlands ; organized by UDC Consortium, The Hague*. Würzburg : Ergon. ISBN 9783956500077 3956500075.

KESSOUS, Emmanuel, 2012. *L'attention au monde: Sociologie des données personnelles à l'ère numérique*. Armand Colin. ISBN 9782200286729

LA BARRE, Kathryn. « Sempre avanti ? Some reflections on faceted interfaces », dans INTERNATIONAL UDC SEMINAR, SLAVIC, 2013.

LAVOIE, Brian F., SCHONFELD, Roger C. «Books without Boundaries : A Brief Tour of the System-wide Print Book Collection » dans DEMPSEY, Lorcan, LAVOIE, Brian F., MALPAS, Constance, CONNAWAY, Lynn S., SCHONFELD, Roger C., SHIPENGROVER J.D. et WAIBEL, Günter. 2013. *Understanding the Collective Collection : Towards a System-wide Perspective on Library Print Collections*. Dublin, Ohio : OCLC Research. Consulté le 5 août 2014. Disponible à l'adresse Web : http://oclc.org/research/publications/library/2013/2013-09.pdf.

LIN, Xia, AHN, Jae-WOOK. « Challenges of knowledge structure visualization », dans INTENRATIONAL UDC SEMINAR, SLAVIC, 2013.

O'NEIL, Cathy, SCHUTT, Rachel. *Doing Data Science*, [sans date]. [en ligne]. [Consulté le 1 novembre 2014]. Disponible à l'adresse : http://shop.oreilly.com/product/0636920028529.do

POISSENOT, Claude. « La connaissance des publics via les données internes de la bibliothèque. » dans EVANS. 2011.

RAZPOTNIK, Špela, ŠAUPERL, Alenka. « Enhancing browsing experience through visual presentation of subject terms », dans INTERNATIONAL UDC SEMINAR, SLAVIC, 2013.

TESNIÈRE, Valérie, dans BERTRAND, Anne-Marie, BETTEGA, Emilie, CLÉMENT, Catherine, ERMAKOFF, Thierry, EVANS, Christophe, ION, Christina, PICARD, David-Georges, RAPATEL, Livia, TESNIÈRE, Valérie. *Quel modèle de bibliothèque?*, 2008. Presses de l'ENSSIB. ISBN 9782910227739.

THOMAS, Neal. 2012. « Algorithmic subjectivity and the need to be in-formed. » dans LATZKO-TOTH, Guillaume, MILLERAND, Florence. *TEM 2012 : Proceedings of the Technology & Emerging Media Track – Annual Conference of the Canadian Communication Association (Waterloo, May 30 D June 1, 2012).* Consulté le 3 août 2014. http://www.tem.fl.ulaval.ca/www/wpcontent/PDF/Waterloo_2012/THOMASFTEM2012.pdf

TILLICH, Paul et GOUNELLE, André, 2012. *Dynamique de la foi*. Genève; Québec; [Paris] : Éd. Labor et fides ; les Presses de l'Université Laval ; [diff. les Éd. du Cerf]. ISBN 9782830914801 2830914805 9782763796024 2763796028.

TUFTE, Edward. 2001. *The Visual Display of Quantitative Information*, "Graphical Excellence." Cheshire, Connecticut: Graphics Press.

WEINBERGER, David. 2008. *Everything Is Miscellaneous: The Power of the New Digital Disorder*. Henry Holt and Company. ISBN 9780805088113.

WEINGART, Scott B. « From trees to webs : uprooting knowledge through visualization » dans NTERNATIONAL UDC SEMINAR, SLAVIC. 2013.

WRIGHT, Alex, 2008. *Glut: Mastering Information Through the Ages*. Cornell University Press. ISBN 0801475090.

YAU, Nathan, 2013. *Data visualisation: De l'extraction des données à leur représentation graphique*. Editions Eyrolles. ISBN 9782212135992.

Revues

ALAIN, Corbin, 1991. Arlette Farge, « Le goût de l'archive ». *Annales. Économies, Sociétés, Civilisations*. 1991. Vol. 46, n° 3, p. 595-597.

DARNTON, Robert. [sans date]. La chandelle de Jefferson. *Le débat*. [en ligne]. [Consulté le 7 août 2014]. Disponible à l'adresse : http://le-debat.gallimard.fr/articles/2012-3-la-chandelle-de-jefferson/

CARDON, Dominique, 2013. Dans l'esprit du PageRank. *Réseaux*. 1 avril 2013. Vol. 177, n° 1, pp. 63-95. DOI 10.3917/res.177.0063.

DENNI, Gaëlle, 2010. Quatre catégories d'outils pour l'auto-évaluation au SICD2 de Grenoble. [en ligne]. 1 janvier 2010. [Consulté le 26 juillet 2014]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2010-04-0023-005

DRUCKER, Johanna, 2010. 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.

DRUCKER, Johanna, 2011. Humanities Approaches to Graphical Display. [en ligne]. 2011. Vol. 5, n° 1. [Consulté le 1 novembre 2014]. Disponible à l'adresse : http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

E.LINK, Forrest, TOSAKA, Yuji, WENG, Cathy. « Mining and Analyzing Circulation and ILL Data for Informed Collection Devlopment. » Preprint à paraître dans *College & Research Libraries*, 2015. Microsoft Word - Link-Tosaka-Weng.docx - crl14-632.full.pdf, [sans date]. [en ligne]. [Consulté le 8 décembre 2014]. Disponible à l'adresse : http://crl.acrl.org/content/early/2014/10/20/crl14-632.full.pdf

ERDMANN, Christopher, 2014. Teaching librarians to be data scientists. *Information outlook* [en ligne]. mai-juin 2014. Vol. 18, n° 3. [Consulté le 17 août 2014]. DOI 10.5281/zenodo.11217. Disponible à l'adresse : https://zenodo.org/record/11217/files/DataScientistTraining.pdf

GILLESPIE, Tarleton. « The relevance of algorithms », à paraître dans GILLEPSIE, Tarleton, BOCZCOWSKI, Pablo et KIRSTEN, Foot. *Media Technologies*. Cambridge, MA : MIT Press. Consulté le 3 août 2014 à l'adresse Web : http://www.tarletongillespie.org/essays/Gillespie%20-%20The%20Relevance%20of%20Algorithms.pdf.

POULAIN, Martine, 2002. Retourner à Tocqueville. [en ligne]. 1 janvier 2002. [Consulté le 11 décembre 2014]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2002-05-0066-001/2002/5/fam-apropos/varia

ROUVROY, Antoinette et BERNS, Thomas, 2013. Gouvernementalité algorithmique et perspectives d'émancipation. *Réseaux*. 1 avril 2013. Vol. 177, n° 1, pp. 163-196. DOI 10.3917/res.177.0163. p. 180.

« The End of Theory: The Data Deluge Makes the Scientific Method Obsolete ». *WIRED*. Consulté le 2 août 2014. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

TULEU, Benoît, 2011. Trop loin, trop proche. [en ligne]. 1 janvier 2011. [Consulté le 11 décembre 2014]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2011-02-0014-002

SITES INTERNET

Amazon.fr : Hamlet, [sans date]. [en ligne]. [Consulté le 8 septembre 2014]. Disponible à l'adresse : http://www.amazon.fr/s/ref=nb_sb_noss_1?__mk_fr_FR=%C3%85M%C3%85%C5%BD%C3%95%C3%91&url=search-alias%3Daps&field-keywords=Hamlet

About | metaLAB (at) Harvard, [sans date]. [en ligne]. [Consulté le 7 août 2014]. Disponible à l'adresse : http://metalab.harvard.edu/about/

ADMIN, 2012. Data Mining Research Area. [en ligne]. 4 août 2012. [Consulté le 29 janvier 2014]. Disponible à l'adresse : http://oclc.org/research/activities/mining.html

ANDERSON, Rick [sans date]. What Patron-Driven Acquisition (PDA) Does and Doesn't Mean: An FAQ. *The Scholarly Kitchen* [en ligne]. [Consulté le 6 décembre 2014]. Disponible à l'adresse : http://scholarlykitchen.sspnet.org/2011/05/31/what-patron-driven-acquisition-pda-does-and-doesnt-mean-an-faq/

Astronomy Texts in the Internet Archive, [sans date]. *Tableau Software* [en ligne]. [Consulté le 21 août 2014]. Disponible à l'adresse : http://public.tableausoftware.com/views/AstronomyTextsintheInternetArchive/Whatwasthetopdownloadedastronomywork?:showVizHome=no

CAVALIÉ, Etienne, [sans date]. Mais que fait Gephi ? *Bibliothèques [reloaded]* [en ligne]. [Consulté le 17 juillet 2014]. Disponible à l'adresse : http://bibliotheques.wordpress.com/2014/07/03/mais-que-fait-gephi/

Choropleth_US_libs_by_county.jpg (Image JPEG, 1017 × 653 pixels) - Redimensionnée (96%), [sans date]. [en ligne]. [Consulté le 21 août 2014]. Disponible à l'adresse : http://hangingtogether.org/wp-content/uploads/2013/07/Choropleth_US_libs_by_county.jpg

COHEN, Dan, 2012. Visualizing the Uniqueness, and Conformity, of Libraries. *Dan Cohen* [en ligne]. 13 décembre 2012. [Consulté le 11 juin 2014]. Disponible à l'adresse : http://www.dancohen.org/2012/12/13/visualizing-the-uniqueness-and-conformity-of-libraries/

Content Management Services for Libraries and Publishers, [sans date]. [en ligne]. [Consulté le 8 décembre 2014]. Disponible à l'adresse : http://www.swets.fr/

Data Mining « Big Data »: A Strategy for Improving Library Discovery | Blog | Serials Solutions, [sans date]. [en ligne]. [Consulté le 9 mai 2014]. Disponible à l'adresse : http://www.serialssolutions.com/en/words/detail/data-mining-big-data-a-strategy-for-improving-library-discovery.

Dissertation Browser | Information, [sans date]. [en ligne]. [Consulté le 23 mai 2014]. Disponible à l'adresse : http://www-nlp.stanford.edu/projects/dissertations/

Du jeu de données au jeu avec les données | The Pixel Hunt, [sans date]. [en ligne]. [Consulté le 30 juillet 2014]. Disponible à l'adresse : http://florentmaurin.com/?p=471

DST4L Class Notes - Google Docs, [sans date]. [en ligne]. [Consulté le 26 juillet 2014]. Disponible à l'adresse : https://docs.google.com/document/d/1WUz4UwwRv5szcsODIwcEV7qAGNc0gjL-oDErFQ2MoBY/edit?pli=1

FRANCE, Bibliothèque nationale de, [sans date]. BnF - Les enjeux du web de données en bibliothèque. [en ligne]. [Consulté le 2 novembre 2014]. Disponible à l'adresse : http://www.bnf.fr/fr/professionnels/innov_num_web_donnees/a.web_donnees_enjeux_bibliotheques.html

grapheprc3aats.png (Image PNG, 1024 × 1024 pixels) [sans date]. [en ligne]. [Consulté le 20 août 2014]. Disponible à l'adresse : https://bibliotheques.files.wordpress.com/2014/07/grapheprc3aats.png.

GULLIGAN, Finbar. Sans date. Patron-driven library - Patron-driven acquisition - Research Information. [en ligne]. [Consulté le 3 décembre 2014].

HARRIS, Jonathan, KAMVAR, Sep. [sans date]. We Feel Fine. [en ligne]. [Consulté le 20 août 2014]. Disponible à l'adresse : http://wefeelfine.org/

HICKEY, Thomas B., TOVES, Jenny. 2009. « FRBR Work-Set Algorithm, v. 2.0 ». OH: OCLC Online Computer Library Center, Inc. (Research division). Consulté le 4 août 2014 à l'adresse Web : http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf

How to Beat Bibliographic Data into Submission, pt. 1 | Data Scientist Training for Librarians, [sans date]. [en ligne]. [Consulté le 7 juillet 2014]. Disponible à l'adresse : http://altbibl.io/dst4l/how-to-beat-bibliographic-data-into-submission-pt-1/

How to Beat Bibliographic Data into Submission, pt. 2 | Data Scientist Training for Librarians, [sans date]. [en ligne]. [Consulté le 7 juillet 2014]. Disponible à l'adresse : http://altbibl.io/dst4l/how-to-beat-bibliographic-data-into-submission-pt-2/

Internet Archive Search Engine. [Sans date]. Consulté le 19 août 2014. Disponible à l'adresse Web : http://archive.org/advancedsearch.php#raw

Lev Manovich – What is Visualization? | Data Visualisation, [sans date]. [en ligne]. [Consulté le 30 juin 2014]. Disponible à l'adresse : http://www.datavisualisation.org/2010/11/lev-manovich-what-is-visualization/

Library Observatory, [sans date]. [en ligne]. [Consulté le 29 janvier 2014]. Disponible à l'adresse : http://www.libraryobservatory.org/

Library Data Visualization, [sans date]. [en ligne]. [Consulté le 20 mai 2014]. Disponible à l'adresse : http://librarydatavisual.blogspot.fr/

LOUKISSAS, Yanni, [sans date]. Data Artifacts Rising: Cultures of Collecting from Preservation to Participation | metaLAB (at) Harvard. [en ligne]. [Consulté le 19 mai 2014]. Disponible à l'adresse : http://metalab.harvard.edu/2012/12/data-artifacts-rising-cultures-of-collecting-from-preservation-to-participation/

MALPAS, Constance. [sans date]. Sliding scale: mapping local, group and system-wide library infrastructure | hangingtogether.org. [en ligne]. [Consulté le 21 juillet 2014]. Disponible à l'adresse : http://hangingtogether.org/?p=3149

mbattles_udcseminar2013.pdf, [sans date]. [en ligne]. [Consulté le 1 septembre 2014]. Disponible à l'adresse : http://www.udcds.com/seminar/2013/media/slides/mbattles_udcseminar2013.pdf

Penn Library Data Farm, [sans date]. [en ligne]. [Consulté le 13 mai 2014]. Disponible à l'adresse : http://datafarm.library.upenn.edu/

Penn Library - Graduate Student Workshops, [sans date]. [en ligne]. [Consulté le 16 août 2014]. Disponible à l'adresse : http://datafarm.library.upenn.edu/desksurvey/index.html

PRENTICE, Jennifer, ALSTINE, Colin Van, BENSON, Amy et FORD, Jacqueline, 2013. *ADS Monograph Matches in the Internet Archive (Excel)* [en ligne]. juin 2013. [Consulté le 19 août 2014]. Disponible à l'adresse : http://figshare.com/articles/ADS_Monograph_Matches_in_the_Internet_Archive/710921

SAO/NASA ADS Custom Query Form, [sans date]. [en ligne]. [Consulté le 19 août 2014]. Disponible à l'adresse : http://adsabs.harvard.edu/abstract_service.html

Sizing Up Big Data, Broadening Beyond the Internet, [sans date]. *Bits Blog* [en ligne]. [Consulté le 1 novembre 2014]. Disponible à l'adresse : http://bits.blogs.nytimes.com/2013/06/19/sizing-up-big-data-broadening-beyond-the-internet/.

Sliding scale: mapping local, group and system-wide library infrastructure | hangingtogether.org, [sans date]. [en ligne]. [Consulté le 21 juillet 2014]. Disponible à l'adresse : http://hangingtogether.org/?p=3149

« su:Hamlet (Legendary character) Drama ». [WorldCat.org], [sans date]. [en ligne]. [Consulté le 8 septembre 2014]. Disponible à l'adresse : http://www.worldcat.org/search?q=su%3AHamlet+%28Legendary+character%29+Drama.&qt=hot_subject

Tag Galaxy, [sans date]. [en ligne]. [Consulté le 9 septembre 2014]. Disponible à l'adresse : http://taggalaxy.de/

The Life and Death of Data, [sans date]. [en ligne]. [Consulté le 2 novembre 2014]. Disponible à l'adresse : http://lifeanddeathofdata.org/

Top-250-CIC-borrowers-by-location.jpg (Image JPEG, 658 × 435 pixels), [sans date]. [en ligne]. [Consulté le 21 août 2014]. Disponible à l'adresse : http://hangingtogether.org/wp-content/uploads/2013/07/Top-250-CIC-borrowers-by-location.jpg.

UDC Seminar 2013, [sans date]. [en ligne]. [Consulté le 16 mai 2014]. Disponible à l'adresse : http://seminar.udcc.org/2013/programme.php

Visualizing Network Flows: Library Inter-lending | hangingtogether.org, [sans date]. [en ligne]. [Consulté le 3 juin 2014]. Disponible à l'adresse : http://hangingtogether.org/?p=3053

VuFind FAQ: Frequently Asked Questions, [sans date]. [en ligne]. [Consulté le 29 août 2014]. Disponible à l'adresse : http://www.library.illinois.edu/learn/find/vufind/vufind_faq.html.

Watch_Dogs WeAreData, [sans date]. *Watch_Dogs WeAreData* [en ligne]. [Consulté le 31 août 2014]. Disponible à l'adresse : http://wearedata.watchdogs.com/

What is Summon? | University Libraries | Virginia Tech, [sans date]. [en ligne]. [Consulté le 2 août 2014]. Disponible à l'adresse : http://www.lib.vt.edu/help/summon/what-is-summon.html

xlin_udcseminar2013.pdf, [sans date]. [en ligne]. [Consulté le 8 septembre 2014]. Disponible à l'adresse : http://www.udcds.com/seminar/2013/media/slides/xlin_udcseminar2013.pdf

10 principes fondamentaux – Société – Google, [sans date]. [en ligne]. [Consulté le 11 décembre 2014]. Disponible à l'adresse : http://www.google.fr/intl/fr/about/company/philosophy/

VIDÉOGRAPHIES

Introducing the Knowledge Graph, 2012. [en ligne]. [Consulté le 2 septembre 2014]. Disponible à l'adresse : http://www.youtube.com/watch?v=mmQl6VGvX-c&feature=youtube_gdata_player

Leveraging WorldCat: Data Mining the largest library database in the World, 2013. [en ligne]. [Consulté le 14 juillet 2014]. Disponible à l'adresse : http://www.youtube.com/watch?v=atA2QadzTdY&feature=youtube_gdata_playe

Star Trek: Into Darkness - End Credits: Preview (2013) | SD, [sans date]. [en ligne]. [Consulté le 30 juillet 2014]. Disponible à l'adresse : http://www.youtube.com/watch?v=_-W6XIWOiMA

Tag Galaxy - Create Your Own Flickr Photo Universe, 2011. [en ligne]. [Consulté le 29 août 2014]. Disponible à l'adresse : http://www.youtube.com/watch?v=uDMYByYOCa4&feature=youtube_gdata_player

Table des annexes

ANNEXE 1 : LES SUGGESTIONS EN LIGNE, PLUSIEURS DÉFINITIONS DE L'IDENTITÉ DE L'ŒUVRE.

DOCUMENT 1 : SUGGESTIONS DE WORLDCAT POUR *HAMLET*.[273]

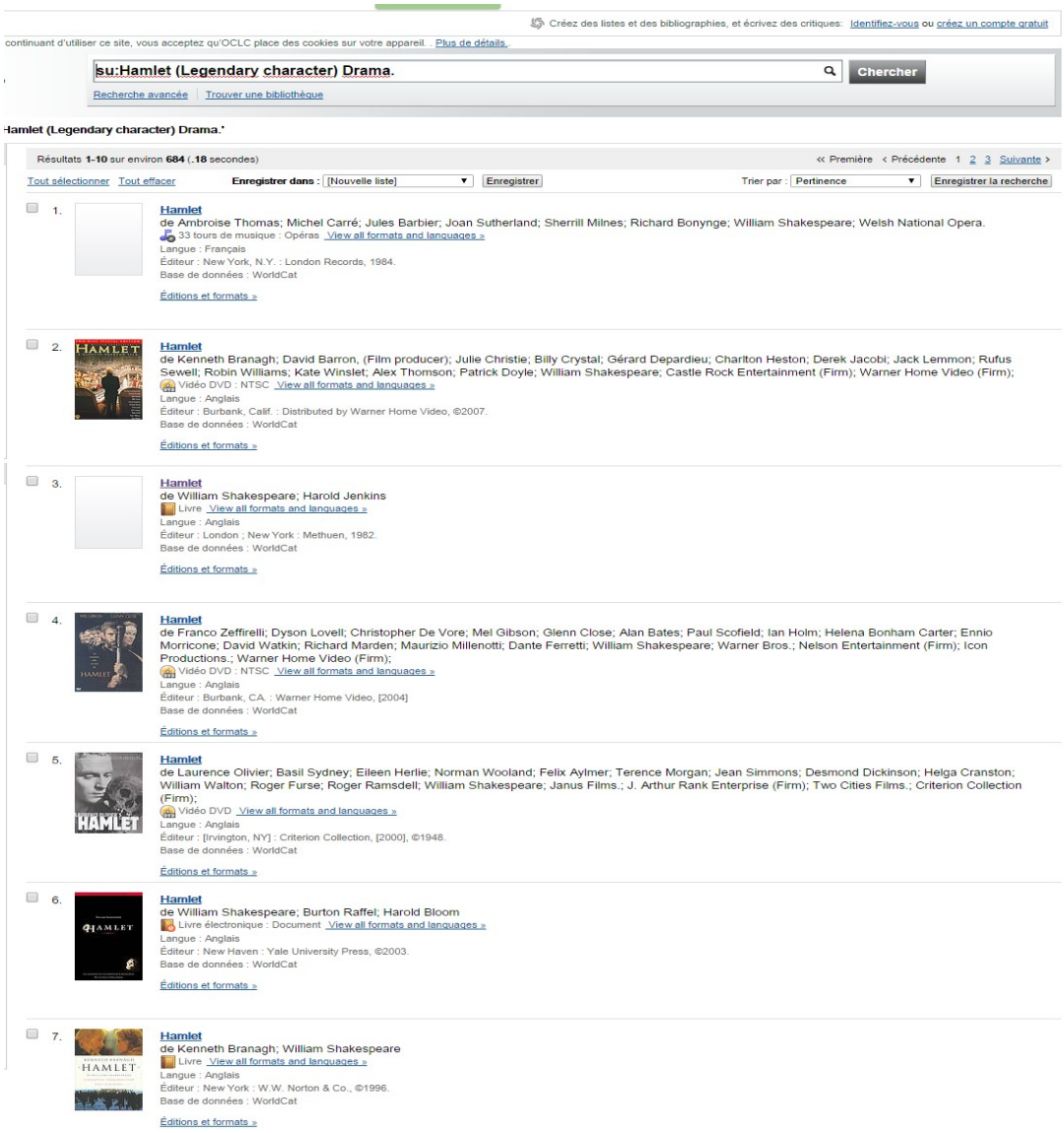


Figure 12 : Suggestions de WorldCat pour Hamlet.

[273] Résultats pour « su:Hamlet (Legendary character) Drama ». [WorldCat.org], [sans date]. [en ligne]. [Consulté le 8 septembre 2014]. Disponible à l'adresse : http://www.worldcat.org/search?q=su%3AHamlet+%28Legendary+character%29+Drama.&qt=hot_subject

DOCUMENT 2 : SUGGESTIONS D'AMAZON POUR HAMLET.[274]

Figure 13 : Suggestions d'Amazon pour Hamlet

[274] Amazon.fr : Hamlet, [sans date]. [en ligne]. [Consulté le 8 septembre 2014]. Disponible à l'adresse : http://www.amazon.fr/s/ref=nb_sb_noss_1?__mk_fr_FR=%C3%85M%C3%85%C5%BD%C3%95%C3%91&url=search-alias%3Daps&field-keywords=Hamlet

ANNEXE 2 : L'OBSERVATOIRE DE LA BIBLIOTHÈQUE.

DOCUMENT 1 : INTERFACE DE L'OBSERVATOIRE, MONTRANT LA TAILLE RELATIVE DES INSTITUTIONS AYANT PARTICIPÉ À LA DPLA.[275]

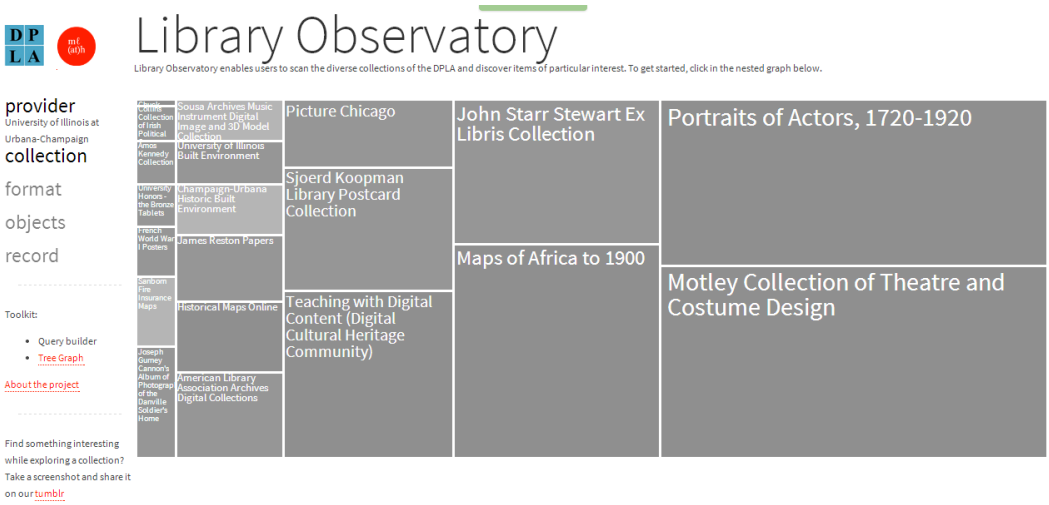


Figure 14 : Visualisation au second niveau de l'institution : Université d'Illinois.

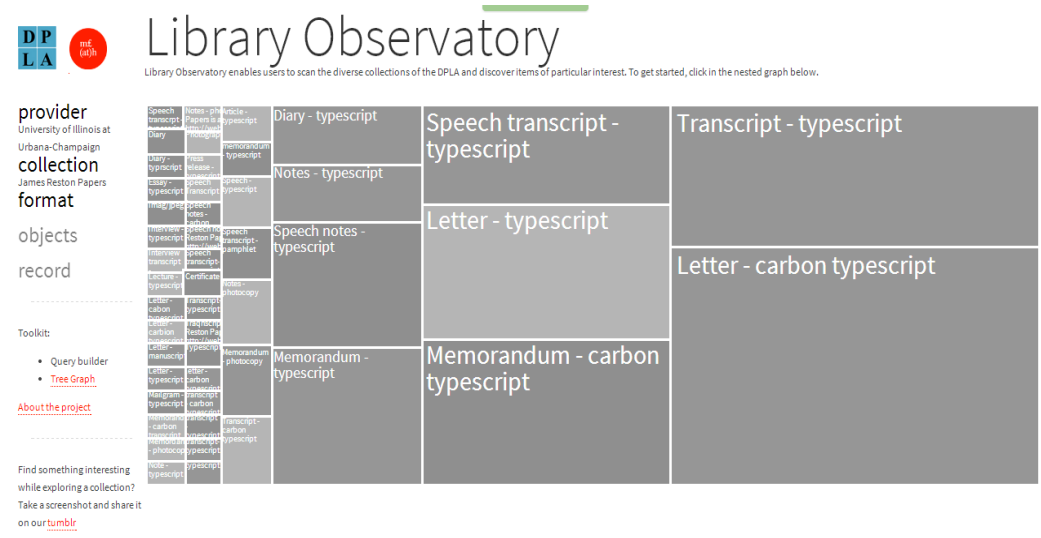


Figure 15 : Visualisation des formats : Université d'Illinois > Collection : James Reston Papers > Format > tapuscrits des notes de discours.

objects

record

Toolkit:

- Query builder
- Tree Graph

About the project

University of Illinois. This resource is provided for research consultation and for scholarly purposes only. Further distribution and/or any commercial use of the works from this site is strictly forbidden without the permission of the University of Illinois and the holder of the copyright in the original manuscript, where copyright in the original manuscript is held by another party. Please contact the University of Illinois Archives at illiarch@uiuc.edu for more information.

link to original record

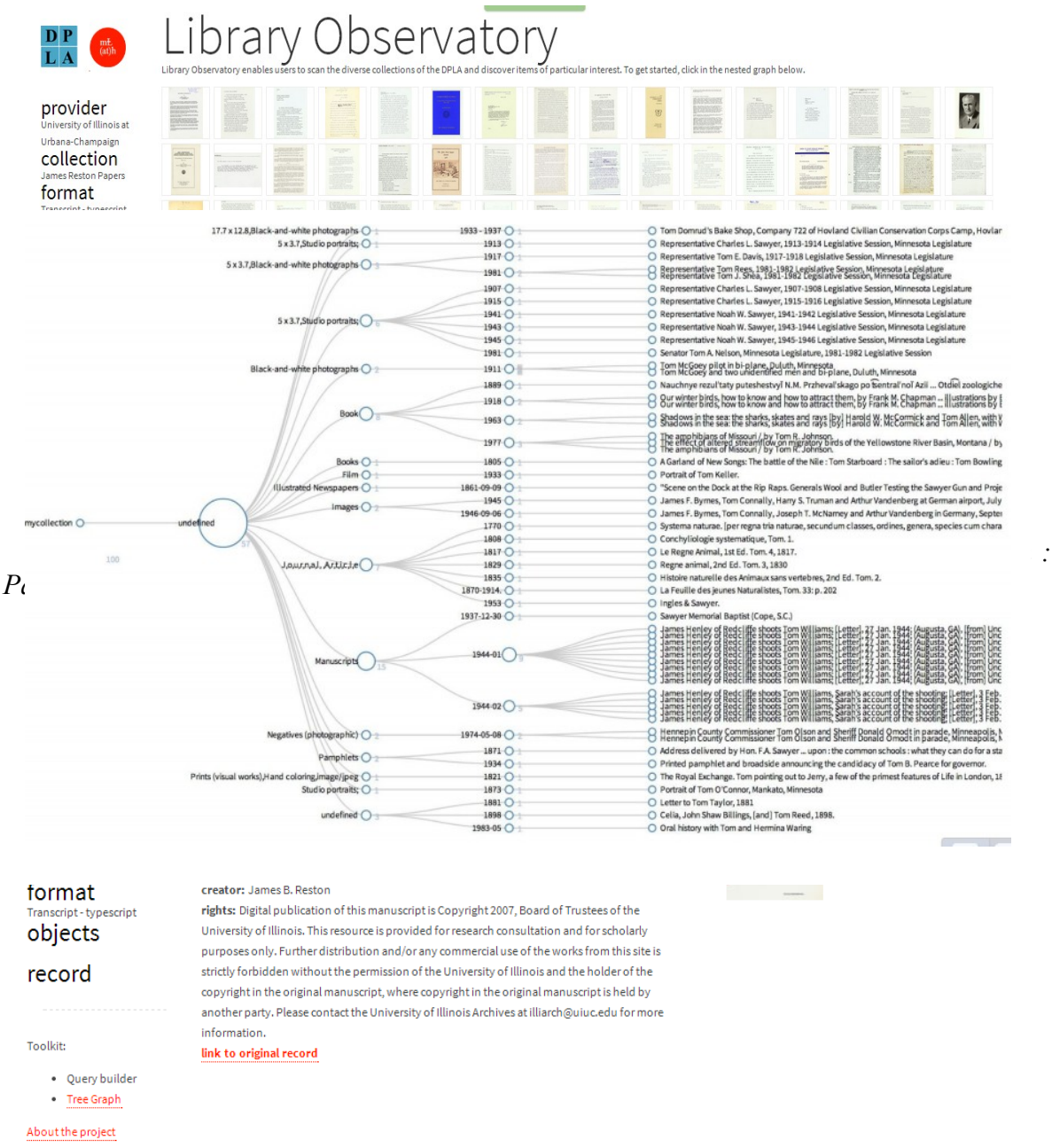


Figure 17 : Visualisation de Document : Université de l'Illinois > Collection : Papiers de James Reston > Format > Tapuscrit de notes de discours > Notes d'un discours prononcé à la cérémonie de remise des diplômes de l'Université de Columbia.

DOCUMENT 2 : VISUALISATION ARBORESCENTE DE LA DPLA.[276]

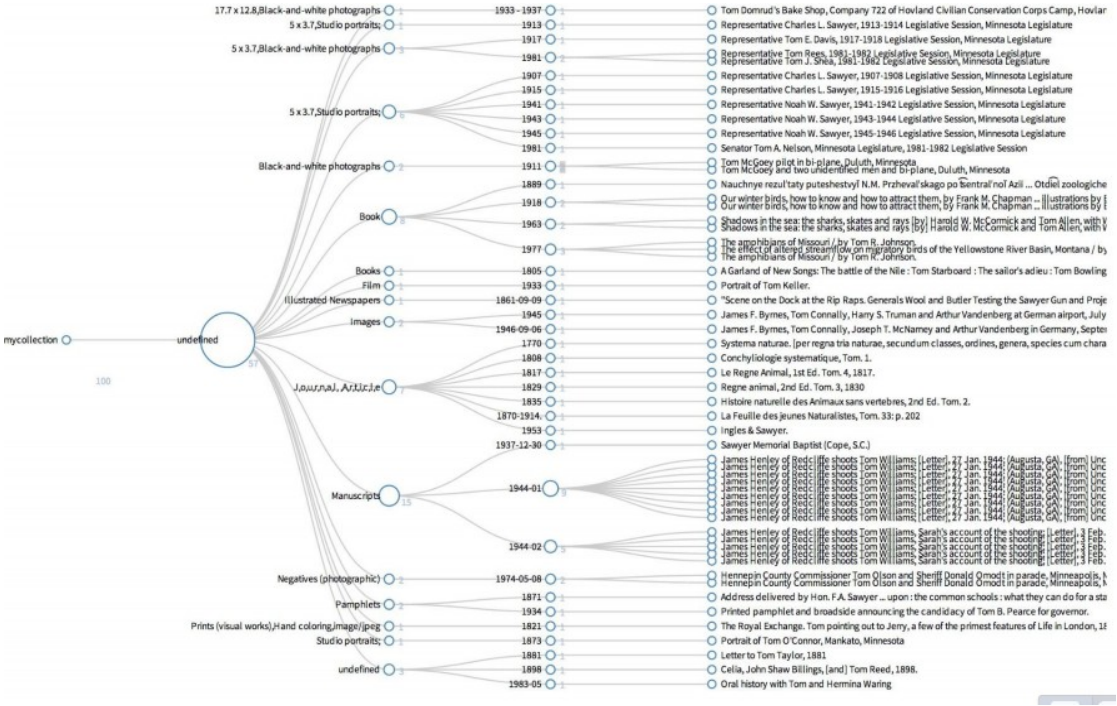


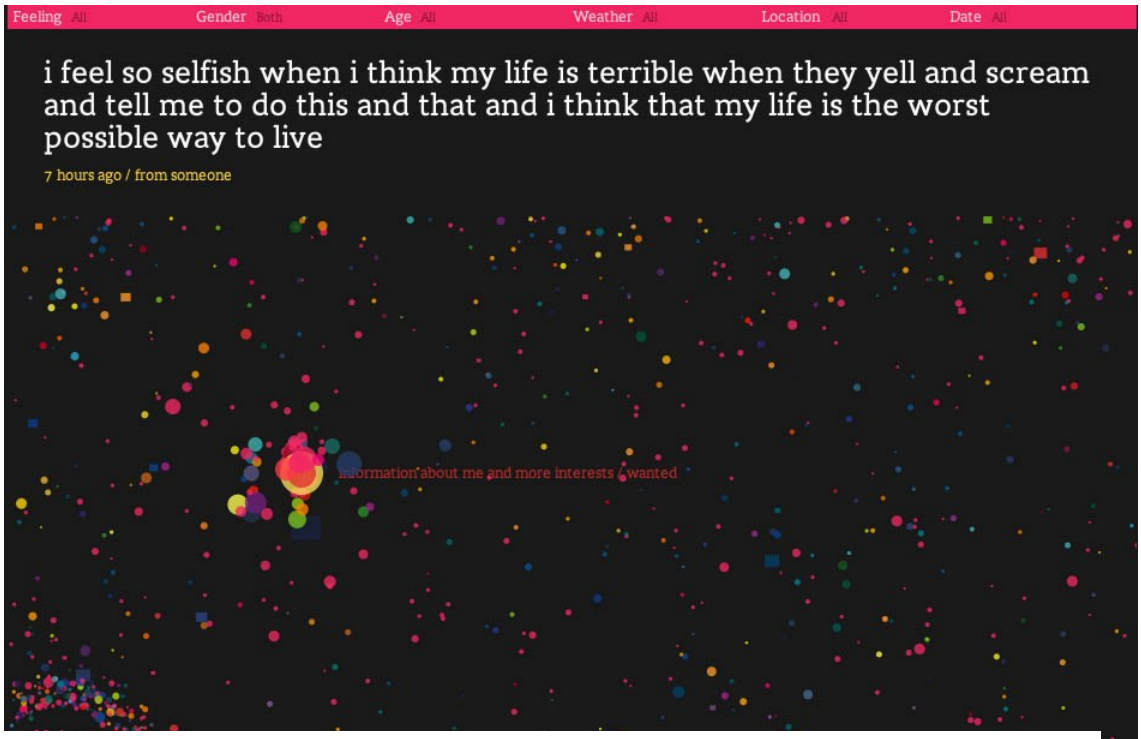
Figure 18 : Visualisation arborescente de la collection de la DPLA.

ANNEXE 3 : LA VISUALISATION AU SERVICE DE LA COMMUNICATION DU BIBLIOTHÉCAIRE VERS SON ÉLU OU

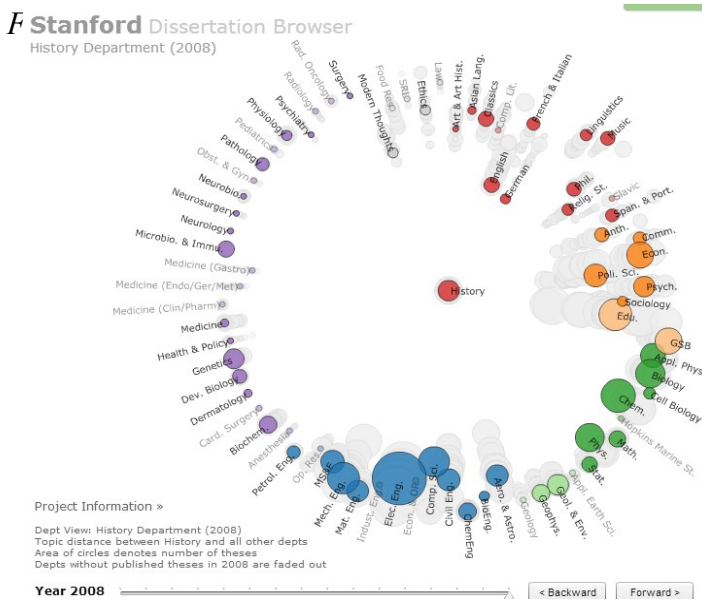
[276] *Ibid*.

SON DIRECTEUR D'UNIVERSITÉ.

DOCUMENT 1 : LA VISUALISATION DES SENTIMENTS : « WE FEEL FINE ».[277]

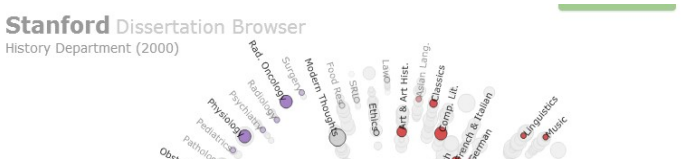


F Stanford Dissertation Browser *nford.*

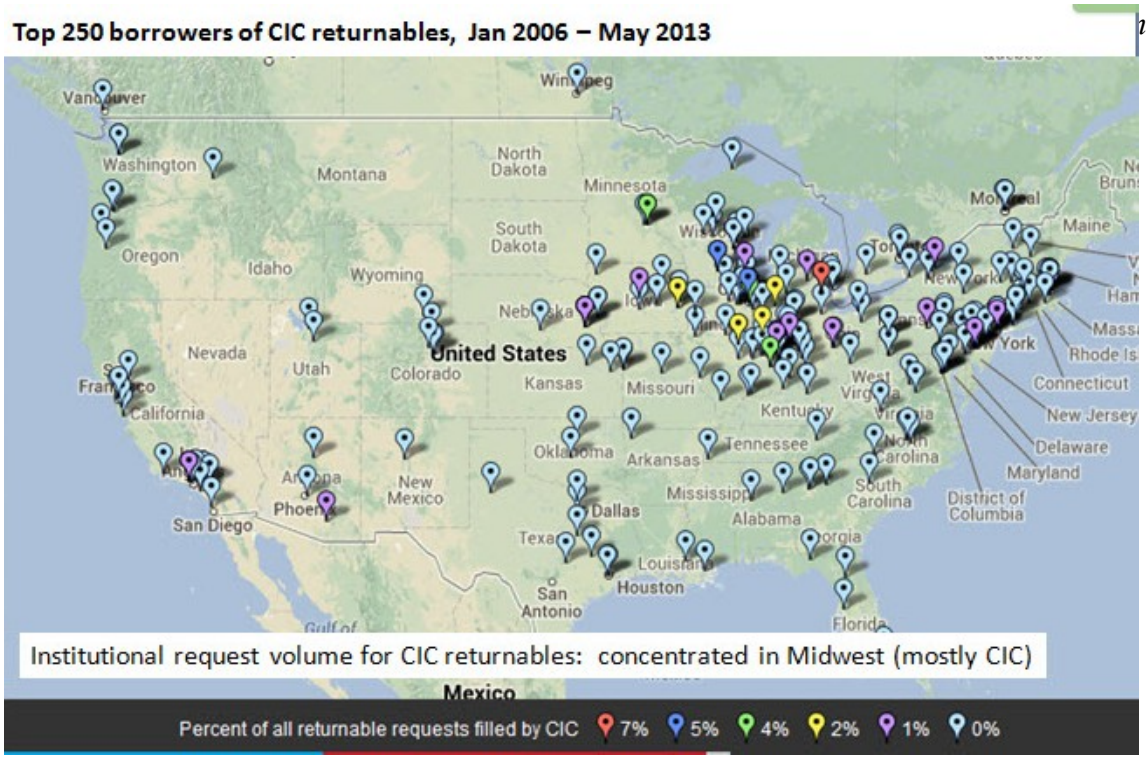


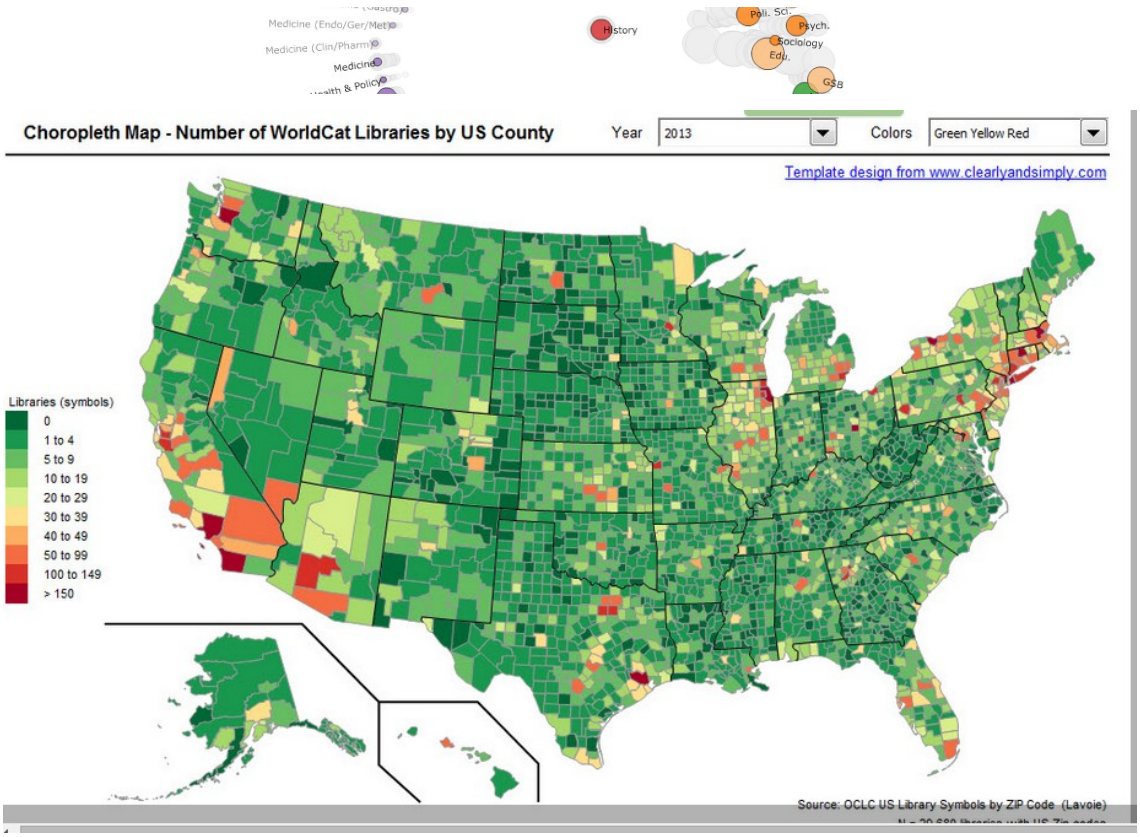
[277] We Feel Fine / by Jonathan Harris and Sep Kamvar, [sans date]. [en ligne]. [Consulté le 20 août 2014]. Disponible à l'adresse : http://wefeelfine.org/. Et We Feel Fine, [sans date]. *Media Molecule* [en ligne]. [Consulté le 8 septembre 2014]. Disponible à l'adresse : http://www.mediamolecule.com/blog/article/we_feel_fine

Document 2 : Le Navigateur de thèses de l'Université de Stanford.[278]



n des





Document 3 : Visualisation des établissements affiliés à WorldCat, dans la perspective du PEB aux États-Unis.[279]

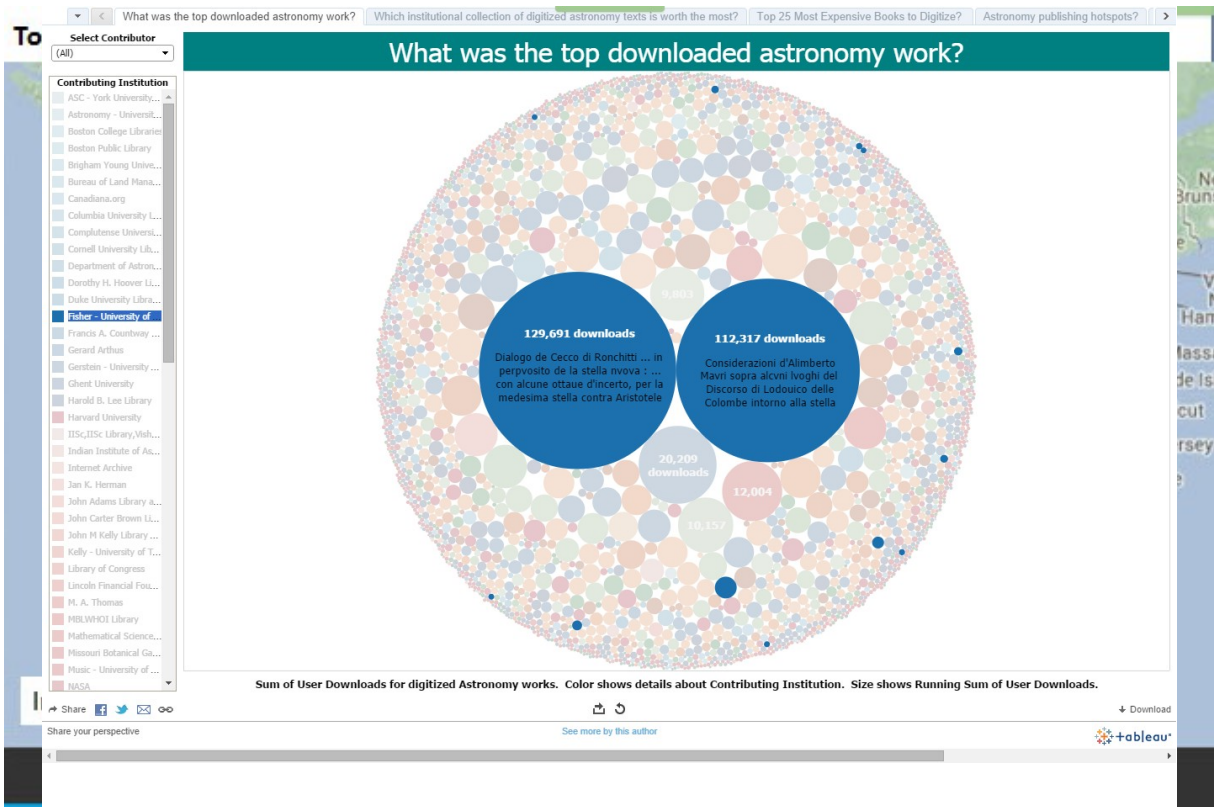


Illustration 23 : Essai de représentation géographique du PEB : visualisation des bibliothèques sous forme de points.

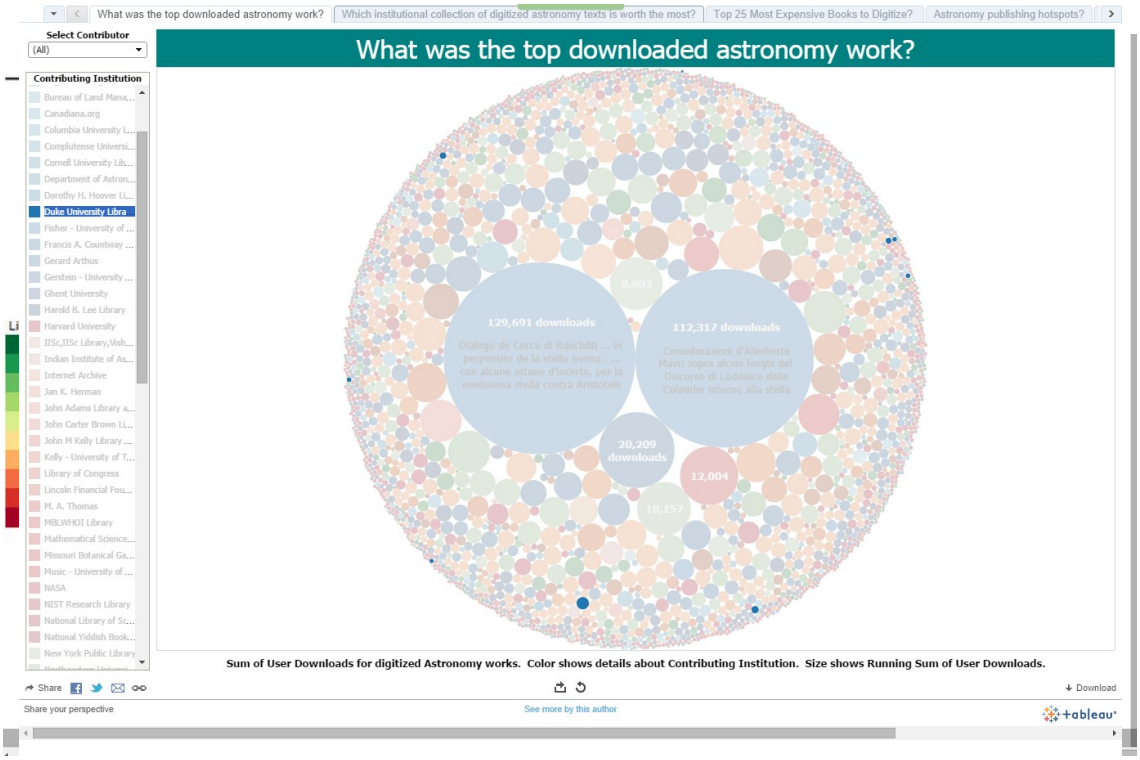


Illustration 23 : Essai de représentation géographique du PEB : visualisation des bibliothèques sous forme de dégradé de couleurs

[279] Sliding scale: mapping local, group and system-wide library infrastructure | hangingtogether.org, [sans date]. [en ligne]. [Consulté le 21 juillet 2014]. Disponible à l'adresse : http://hangingtogether.org/?p=3149

Document 4 : Visualisation des collections d'astronomie présente dans l'Internet Archive par institutions d'origine.[280]

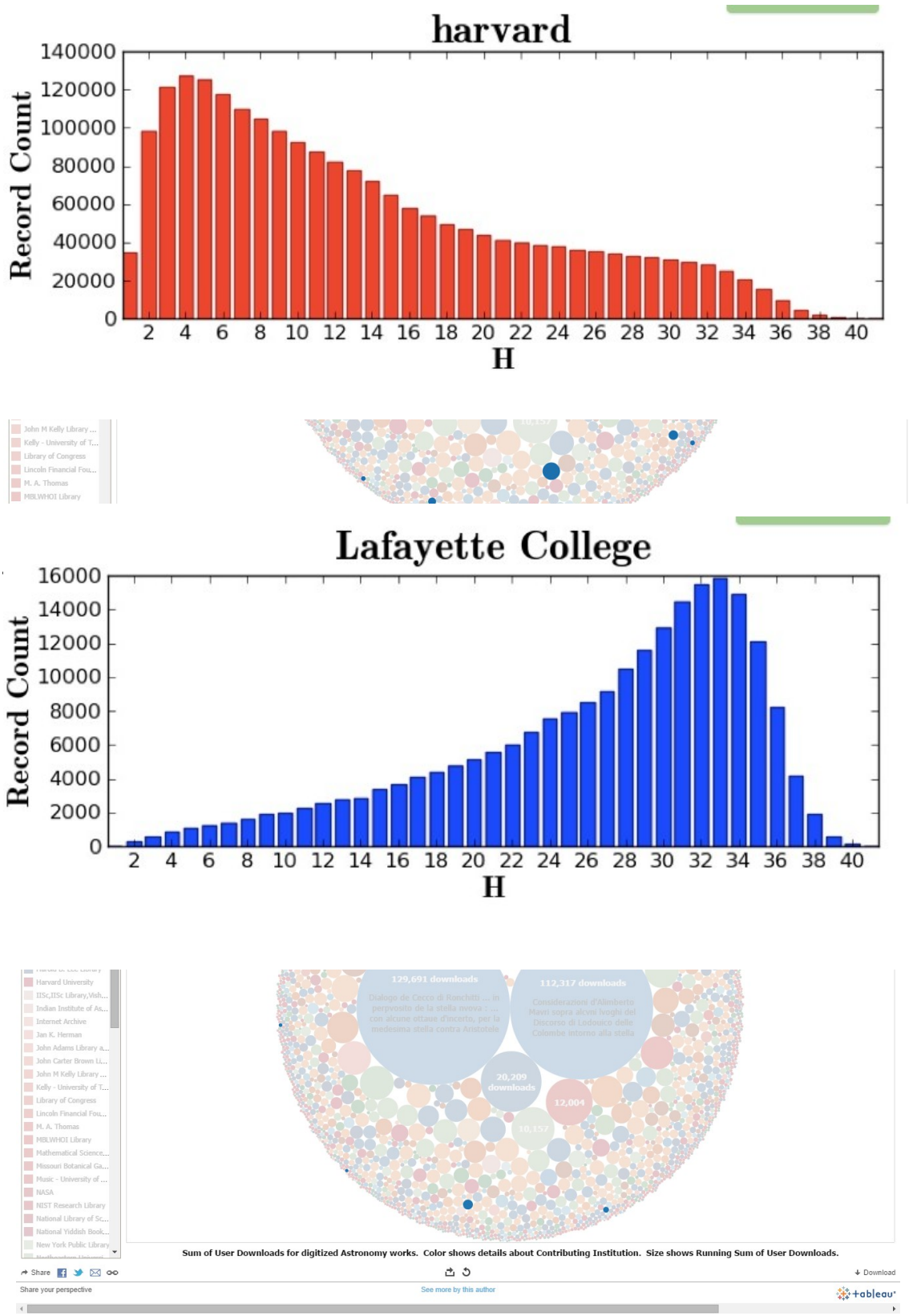


Figure 25 : La collection en Astronomie de la Duke University Library

[280]Astronomy Texts in the Internet Archive, [sans date]. *Tableau Software* [en ligne]. [Consulté le 21 août 2014]. Disponible à l'adresse : http://public.tableausoftware.com/views/AstronomyTextsintheInternetArchive/Whatwasthetopdownloadedastronomywork?:showVizHome=no

Document 4 : Visualisation des Ouvrages les plus détenus à l'échelle globale (ordonnées) et à l'échelle locale (abscisses).[281]

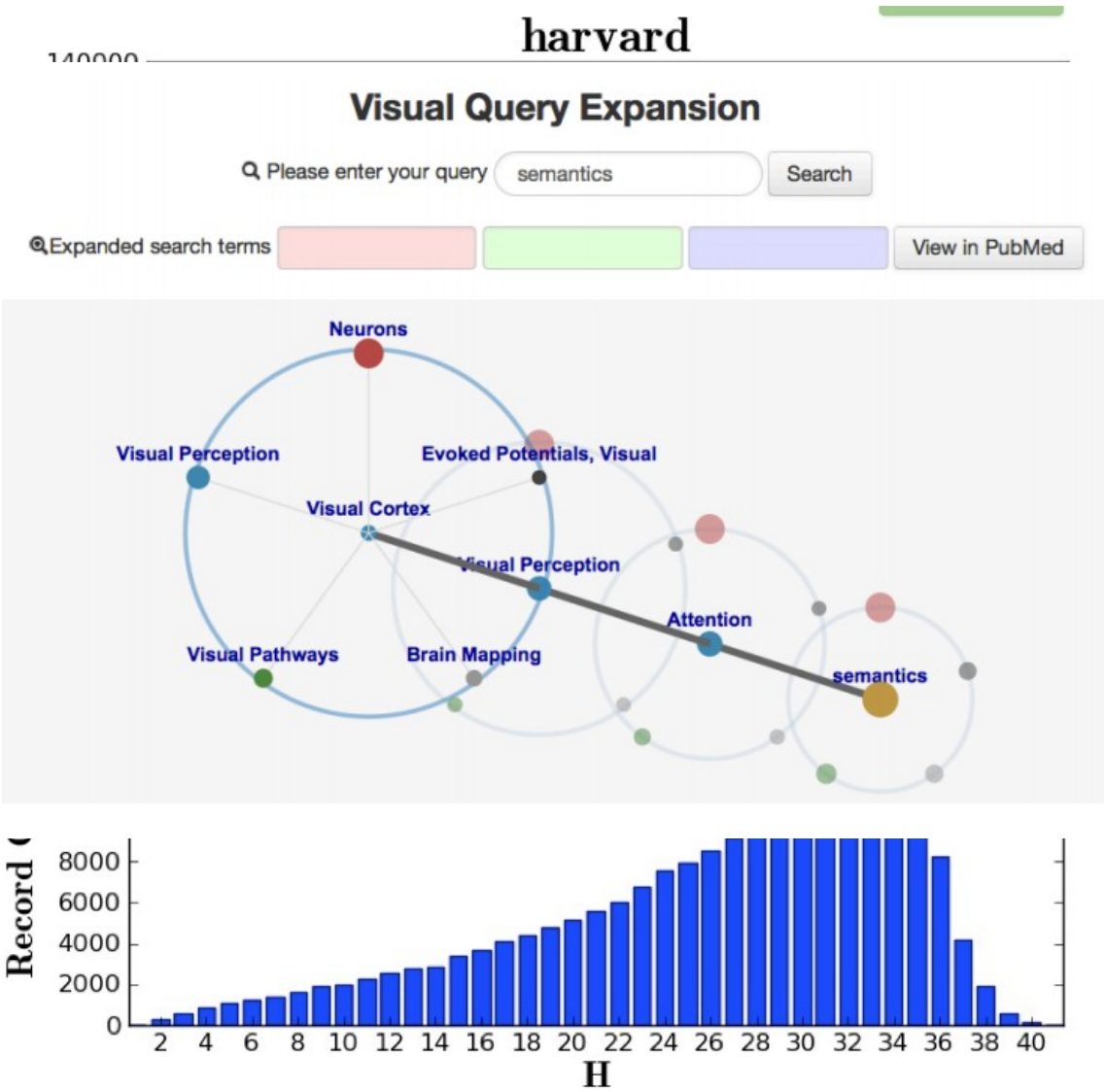


Figure 27 : La collection du Lafayette College, se voulant universelle et grand public.

[281] COHEN, Dan, 2012. Visualizing the Uniqueness, and Conformity, of Libraries. *Dan Cohen* [en ligne]. 13 décembre 2012. [Consulté le 11 juin 2014]. Disponible à l'adresse : http://www.dancohen.org/2012/12/13/visualizing-the-uniqueness-and-conformity-of-libraries/

Figure 29 : Amazon, exemple par excellence d'interface à facettes.

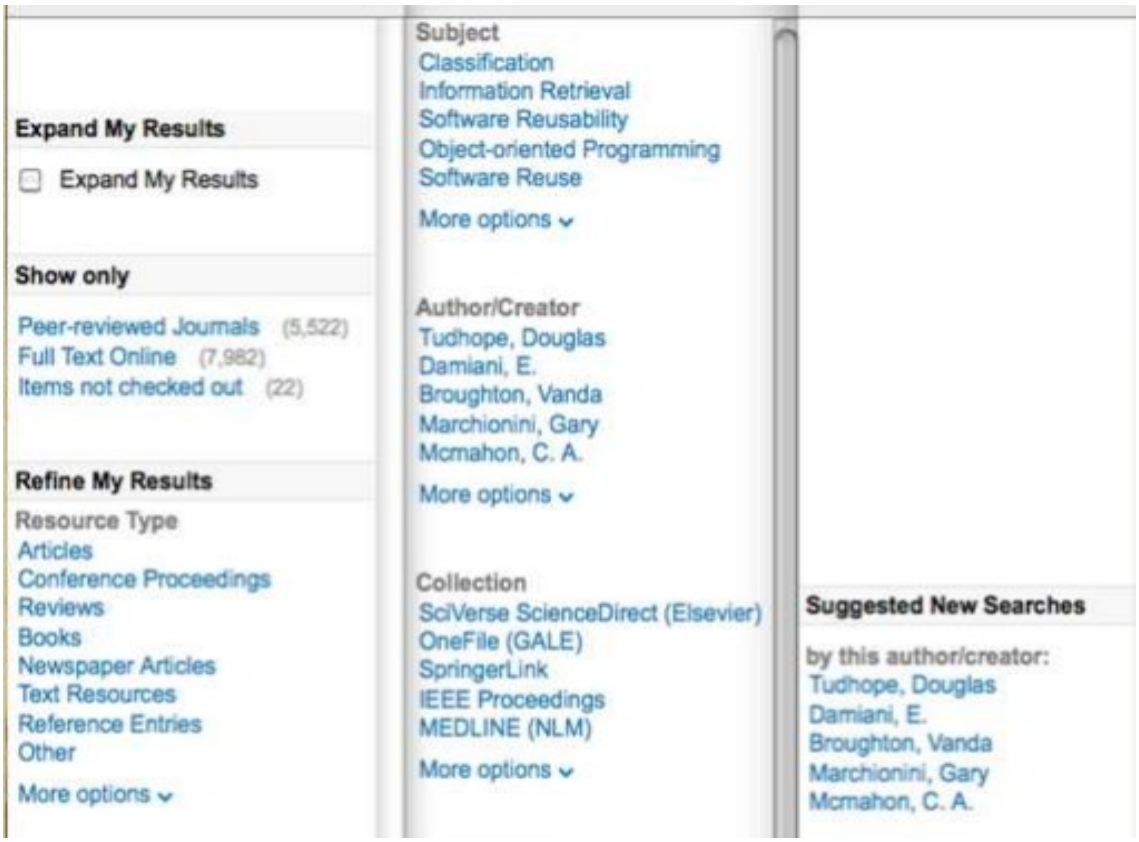
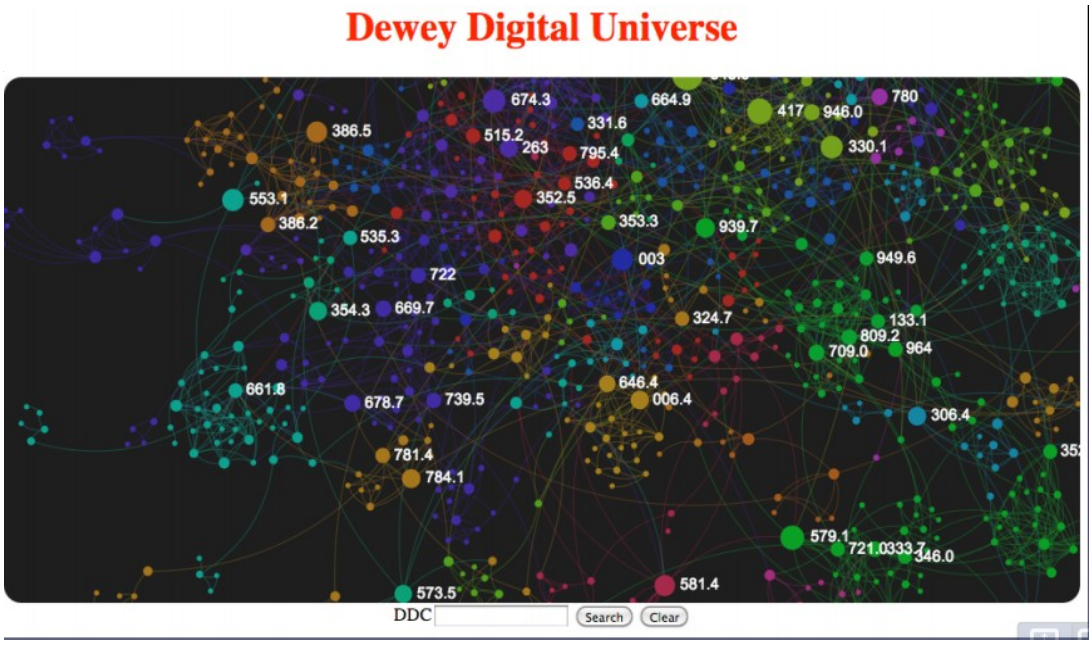


Figure 30 : Affichage Primo de restriction possible en réponse à une requête.

Document 2 : Navigation à facettes.[283]



[283] klabarre_udcseminar2013.pdf, [sans date]. [en ligne]. [Consulté le 8 septembre 2014]. Disponible à l'adresse : http://www.udcds.com/seminar/2013/media/slides/klabarre_udcseminar2013.pdf

Figure 29 : Amazon, exemple par excellence d'interface à facettes.

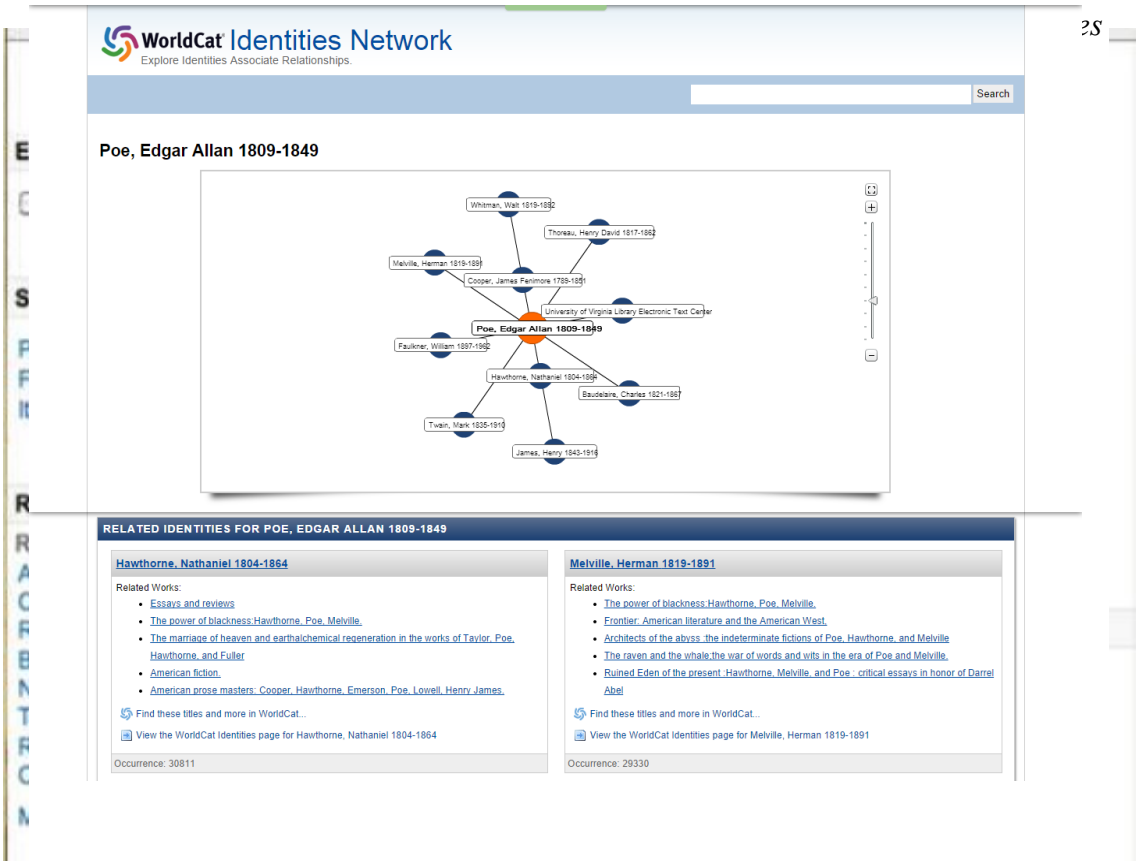


Figure 30 : Affichage Primo de restriction possible en réponse à une requête.

Document 3 : Visualisation sémantique en réseau[284]

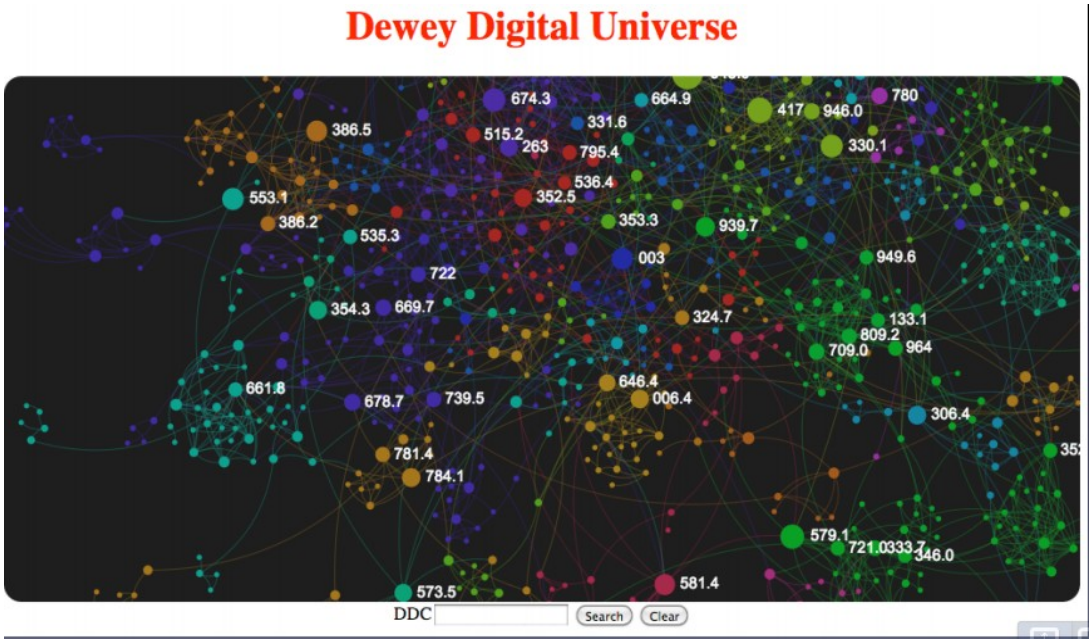


Figure 31 : Visualisation des réseaux sémantiques de la classification Dewey

Document 4 : Visualisation en galaxie[285]

Figure 32 : étoile de photographies sur Flickr.

[284] xlin_udcseminar2013.pdf

[285] Tag Galaxy, [sans date]. [en ligne]. [Consulté le 9 septembre 2014]. Disponible à l'adresse : http://taggalaxy.de/

Document 4 : Un exemple de Data Game.[286]

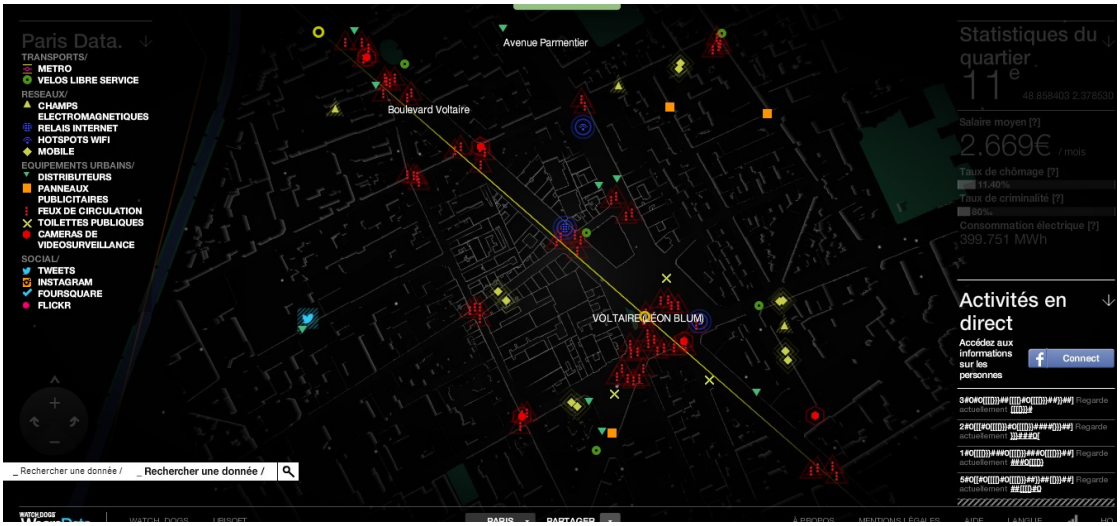


Figure 33 : L'incroyable "We are Data", modélisation interactive des données de Paris, Londres et Berlin. Ici, le onzième arrondissement de Paris.

Document 5 : Une visualisation en réseau pour assister la navigation. [287]

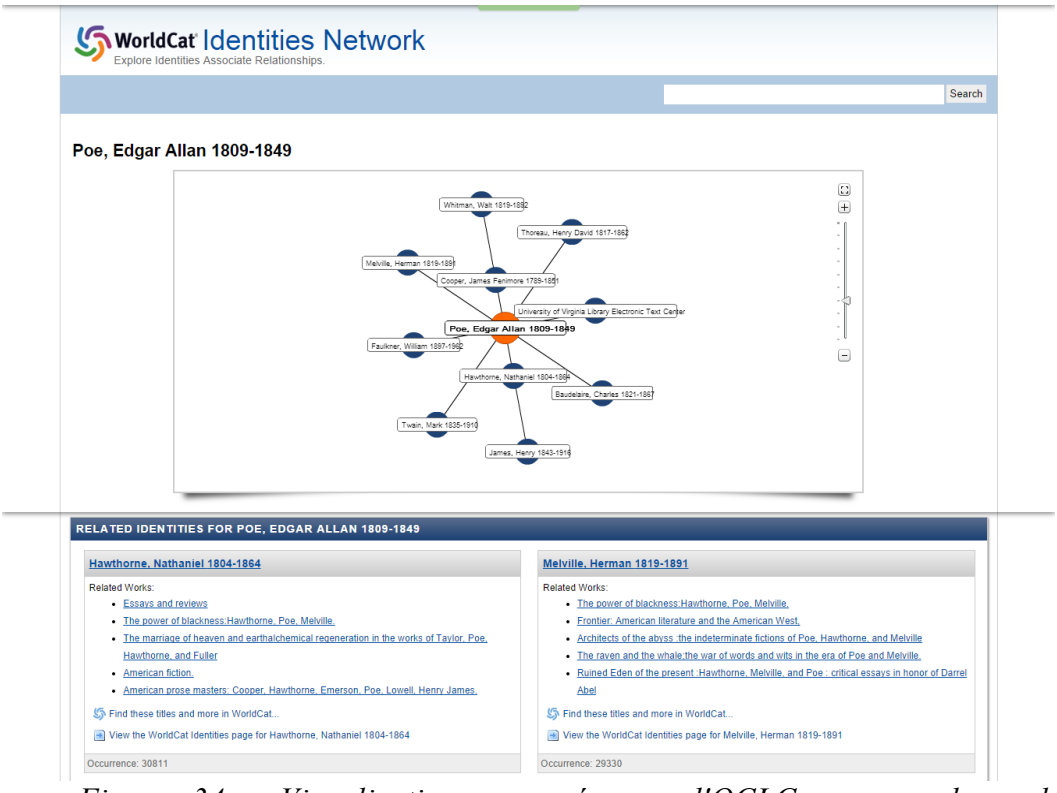


Figure 34 : Visualisation proposée par l'OCLC pour explorer les identités de WorldCat.

[286] Watch_Dogs WeAreData, [sans date]. *Watch_Dogs WeAreData* [en ligne]. [Consulté le 31 août 2014]. Disponible à l'adresse : http://wearedata.watchdogs.com/

[287] RESEARCH, OCLC, [sans date]. WorldCat Identities Network. [en ligne]. [Consulté le 2 septembre 2014]. Disponible à l'adresse : http://experimental.worldcat.org/IDNetwork/display.html?query=lccn-n83162771

Table des matières