Diplôme de conservateur de bibliothèque

Mémoire d'étude / Janvier 2016

Big data et bibliothèques : traitement et analyse informatiques des collections numériques

Johann Gillium

Sous la direction de Monique Joly
Directrice – Service commun de la documentation de l'INSA Lyon

Remerciements

Je souhaite tout d'abord remercier ma directrice de mémoire, Mme Monique Joly, pour ses conseils avisés, les pistes de recherche qu'elle a bien voulu me suggérer tout au long de ce travail, et son attention constante.

Ma gratitude va ensuite à toutes les personnes qui ont bien voulu me recevoir ou répondre à mes questions au cours de la rédaction de ce mémoire : M. Mathieu Andro, ingénieur à l'Institut national de recherche agronomique, M. Pâcome Aurengo, chargé d'études à la Bibliothèque nationale de France, M. Raymond Bérard, directeur de l'INIST-CNRS, Mme Emmanuelle Bermès, conservatrice à la Bibliothèque nationale de France, M. Phillipe Chevalier, responsable des études de la délégation à la stratégie et à la recherche de la BNF, M. Jean-Gabriel Ganascia, directeur-adjoint du Labex Obvil (Observatoire de la vie littéraire), M. Gaël Guibon, ingénieur d'étude chargé du projet d'extraction terminologique de la plate-forme ISTEX, M. Jean-Philippe Moreux conservateur à la Bibliothèque nationale de France, M. Jean-Marie Pierrel, professeur à l'université de Lorraine et chargé de mission de l'Université de Lorraine et de la CPU pour le projet ISTEX, et enfin M. Peter Stirling, chargé de collections numériques au dépôt légal du Web de la BNF.

Si je n'ai pas pu approfondir toutes les pistes de réflexions suggérées au cours des entretiens qui m'ont été accordés, tous ont, d'une manière ou d'une autre, nourri ce travail.

Résumé : ***Cette étude s'attache à présenter sous quels aspects les collections numériques des bibliothèques relèvent des problématiques propres aux données massives, et en quoi les techniques de fouille de données (text and data mining) représentent désormais une nécessité pour l'appropriation par les chercheurs des résultats de la littérature scientifique. Ce travail, qui met au centre de son propos les techniques de fouille de données comme moyens de maîtriser la masse documentaire, identifie trois problématiques distinctes concernant les bibliothèques numériques et ces dispositifs de lecture algorithmiques : sont ainsi abordées successivement les démarches à mettre en œuvre pour aider les chercheurs à faire usage de ces nouvelles méthodes de lecture, puis l'emploi de techniques de fouille de données sur les collections pour constituer de nouvelles formes d'instruments de recherche, et enfin l'usage de la fouille pour assister le traitement documentaire. L'étude se conclut sur le détail des questions juridiques soulevées actuellement par la fouille de données, en rapport avec le droit de la propriété intellectuelle.***

Descripteurs : données massives, exploration de données, bibliothèques virtuelles, humanités digitales, propriété intellectuelle.

Abstract : ***This study aims to display in which aspects libraries's digital collections shares commons features with big data, and how technical devices related to big data, as text and data mining, are now necessary to help searchers to leverage scientific literature. This work revolves around text and data mining, and identifies three differents ways by which TDM can impact digital libraries : first, it becomes necessary to offer services to searchers in order to make easier text and data mining on digital collections. Then, librarians can use TDM in order to build cutting-edge tools for analysing the digital collections. Finally, they can also use TDM to work on digital collections. This study concludes by analysing the current legal issues about text and mining and intellectual property.***

Keywords : big data, text and data mining, scientific literature, digital libraries, digital humanities, computing science, copyright

Droits d'auteurs

Sommaire

Sigles et abréviations

BNF : Bibliothèque nationale de France
CDC : Center for desease control
TDM : Text and data mining
API : Application programming interface
ISTEX : Initiative d'excellence en information scientifique et technique
TALN : Traitement automatique du langage naturel
IFLA : The International Federation of Library Associations and Institutions

INTRODUCTION

À la fois omniprésente et vaporeuse, la notion de mégadonnées [1] s'est répandue cette dernière décennie, gagnant la faveur des médias – à travers des publications remarquées dans des revues prestigieuses, et notamment un numéro spécial de la revue Nature paru en 2008 qui fit date [2] - sans pour autant qu'un consensus se dessine sur ce que signifiait l'expression. Parmi les nombreuses définitions mises en circulation pour caractériser le big data, voici l'une d'entre elles, la plus synthétique et la plus simple pour une entrée en matière : il s'agit d' « ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information »[3]. Encore faut-il immédiatement signaler que cette définition n'épuise pas toutes les significations qu'évoque la notion de big data. À un premier niveau de compréhension [4], on se sert de cette expression comme d'une notion synthétique qui rassemble trois phénomènes : un accroissement du nombre de données produites par l'activité humaine, une amélioration de notre capacité à les stocker, ainsi que le développement de possibilités inédites d'analyse de ces données par des moyens informatiques, notamment en croisant des jeux de données hétérogènes à la recherche de corrélations entre jeux de données. L'ensemble de cette activité ayant pour objectif de déterminer des motifs récurrents, dont on espère qu'ils aideront à la prise de décision en ayant une valeur prédictive.

Bien que les données massives concernent des domaines très divers, et notamment la recherche scientifique, il est juste de dire que le phénomène est fréquemment évoqué dans ses relations avec la sphère marchande : envisagée sous cet éclairage restreint, la notion de « Big data » désigne la collecte et l'analyse des traces numériques de toutes sortes laissées par les individus, le croisement de différents jeux de données à la recherche de corrélation, soit pour proposer des services commerciaux, soit pour les exploiter directement et en tirer des enseignements. Une entreprise emblématique de ce secteur d'activité serait Google. Mais une analyse de données de type « Big data » peut également se faire sans vocation commerciale : le service des pompiers de New-York utilise ainsi un algorithme d'analyse de données massives pour établir quels sont les immeubles les plus susceptibles de présenter un risque d'incendie, en croisant des données telles que les niveaux de revenu des habitants par secteur ou l'ancienneté des immeubles[5].

La notion est donc équivoque, et donne lieu à des réactions contrastées. La crainte d'un futur aliénant marqué par la captation permanentes des traces numériques laissées par individus, leur analyse et leur exploitation aux dépens du respect dû à la vie privée, s'exprime fréquemment sur un ton alarmiste : le big data serait le signe avant-coureur d'une « quantification de soi »[6]. D'autres préfèrent s'attarder avec enthousiasme sur les possibilités que semblent offrir ces nouvelles

[1]Les expressions « mégadonnées », « Big data » et « données massives » seront employées indifféremment dans la suite de ce travail.

[2]*Nature*, Volume 455, N° 7209, 4 septembre 2008. Numéro spécial sous-tiré « Science in the petabyte era ».

[3]« Big data », *Wikipedia* (en ligne : <https://fr.wikipedia.org/wiki/Big_data>) (Consulté le 10 octobre 2014)

[4]Tel qu'un texte de vulgarisation comme le suivant l'expose pages 21 à 23 : Xavier Perret, Guy Jacquemelle. *Big data, le cinéma avait déjà tout imaginé.* Bluffy, Kawa, 2014.

[5] Elizabeth Dwoskin, « How New York's Fire Department Uses Data Mining », *Digits*, 24 janvier 2014 (disponible sur : <http://blogs.wsj.com/digits/2014/01/24/how-new-yorks-fire-department-uses-data-mining/>) (consulté le 29 septembre 2015).

[6]Ces expressions sont employées dans Eric Sadin, *La vie algorithmique : Critique de la raison numérique*, Paris, Éditions L'Échappée, 2013.

méthodes d'analyse, par exemple pour mieux organiser les réponses humanitaires en cas de désastres naturels[7]. Dans le domaine scientifique, certains annoncent, à l'instar du magazine Wired dans un article paru en 2008, l'obsolescence de la méthode scientifique traditionnelle, rendue caduque par les données massives, qui permettent de se passer d'hypothèses ou de modèles théoriques[8]. À entendre l'auteur de cet article, il suffirait désormais d'observer les corrélations entre données pour en tirer pragmatiquement des enseignements, sans plus se préoccuper des modèles théoriques que rendait nécessaires la pénurie de données des temps anciens. Plus sagement, des chercheurs comme l'américaine Danah Boyd rappellent que ces données dont l'agrégation constitue le « Big data » expriment avant tout les préjugés de ceux qui les collectent, et qu'il serait vain d'en attendre des enseignements[9] dépassant leurs biais natifs. Bien que relativement récent, on voit donc que le thème des « données massives » a suscité un intérêt marqué, qui s'exprime à travers d'abondantes prises de position et articles divers, souvent passionnés.

À ce débat, les bibliothécaires ne peuvent rester indifférents, à qui rien de ce qui est informationnel ne devrait être étranger. Aussi bien la question fait-elle l'objet de plus en plus de débats au sein de la profession, et elle a particulièrement été débattue lors du congrès international de l'IFLA qui s'est tenu en 2015 à Cape Town[10], d'abord en ce qui concerne la gestion des données concernant les lecteurs (comme les historiques de prêt) et les enseignements qu'il est possible de tirer de leur analyse. Cette approche est certainement pertinente, mais elle n'embrasse pas le sujet dans toute son étendue, et elle doit être complétée en dressant un tableau d'ensembles des données qu'une bibliothèque produit et gère dans le cadre de son activité. On peut répartir celles-ci en plusieurs catégories : les données relatives à leurs activités internes (acquisition, désherbage, jauges de fréquentation), les données personnelles des usagers, les métadonnées des collections, et, enfin, les collections elles-mêmes. Cette dernière assertion, qui peut sembler surprenante, appelle de plus amples explications. Avant de revenir plus en détail sur le sujet dans une partie ultérieure, rappelons que le texte des collections, pour peu qu'il puisse être disponible sous des formats numériques permettant l'analyse informatique, peut également faire l'objet d'approche de type fouille de texte. Le texte de ces collections, bien qu'il ne relève pas de la « donnée » au sens restreint qu'en donnent d'anciennes définitions[11], peuvent devenir des données par destination, au sens où ce texte peut être analysé par des programmes informatiques qui lui feront subir un ensemble d'opérations, notamment statistiques, qui traiteront le texte comme une donnée de départ.

Après cette entrée en matière, considérons plus attentivement sous quels rapports l'idée de big data peut-elle s'appliquer aux différents types de données

[7]Patrick Meier. *Digital humanitarians : how big data is changing the face of humanitarian response.* Boca raton, CRC Press Taylor & Francis Group, 2015.

[8]Chris Anderson, « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », *Wired*, 23 juin 2008 (disponible en ligne sur : <http://www.wired.com/2008/06/pb-theory/ >) (consulté le 23 novembre 2015).

[9]Hubert Guillaud, « L'inquiétant n'est pas le Big data, c'est qui l'utilise et comment », *Pixels*, 31 octobre 2015 (disponible sur ce site : <http://internetactu.blog.lemonde.fr/2015/10/31/linquietant-nest-pas-le-big-data-cest-qui-lutilise-et-comment/>) (consulté le 2 novembre 2015).

[10]Jean-Philippe Accart, « L'Ifla 2015 comme si vous y étiez : revivez les grands moments du congrès de Cape Town », *Archimag*, 30 octobre 2015 (disponible en ligne sur : <http://www.archimag.com/bibliotheque-edition/2015/10/30/ifla-2015-grands-moments-congres-cape-town>) (consulté le 3 décembre 2015).

[11]Par exemple, Lynda Kellam et Katharin Peter donnaient cette définition de la donnée : « Toute information structurée de manière reconnaissable ». Cité dans Raphaëlle Lapôtre, *Faire parler les données des bibliothèques : du Big data à la visualisation des données*, Villeurbannes, ENSSIB, 2015, p. 12 (disponible en ligne : <http://www.enssib.fr/bibliotheque-numerique/documents/65117-faire-parler-les-donnees-des-bibliotheques-du-big-data-a-la-visualisation-de-donnees.pdf>) (consulté le 3 novembre 2015).

dont la bibliothèque est dépositaire. Dans le cadre de ce mémoire, nous ne pourrons explorer qu'une des voies possibles pour cette réflexion, mais rappelons que notre travail aurait pu prendre des directions bien différentes. À l'occasion d'un premier article synthétique paru sur le sujet, Emmanuelle Bermès a pu proposer une première vue d'ensemble[12] des différents types d'approches expérimentées dans nos établissements, qui relèvent du big data. Tout d'abord, les bibliothèques seront de plus en plus amenées à collecter les données qui servent à élaborer la recherche, tout comme elles collectent depuis longtemps le résultat de cette recherche ; se faisant, elles devront bâtir des architectures propres à conserver ces données, en prenant en compte les caractéristiques d'ensembles qui relèvent de la problématique des données massives. Ensuite, des techniques propres aux mégadonnées peuvent être employées pour l'analyse des données personnelles conservées par les bibliothèques, avec pour objectif de déterminer des motifs récurrents à portée prédictive, d'améliorer l'allocation des ressources pour la politique documentaire. Le seul exemple connu est ici celui du réseau des bibliothèques de Singapour, qui s'est engagé dans une démarche d'analyse des données dont elle disposait sur les prêts par établissement, afin de trouver des motifs récurrents pour lui permettre de mieux prévoir les futurs besoins de ses usagers[13]. Enfin, l'auteure concluait en signalant que de plus en plus les chercheurs étaient désireux d'analyser les collections des bibliothèques dans leurs ensembles.

Ce mémoire s'éloigne des deux premières approches, et il ne sera question ni de données de la recherche, ni des données personnelles des usagers. Il s'agit de travailler sur la notion de mégadonnées dans les rapports qu'elle peut entretenir avec la masse toujours grandissante des collections numériques des bibliothèques. Se demander si cette notion peut être mobilisée avec profit pour travailler sur les fonds numériques, c'est se poser plusieurs questions : avant tout, celle de déterminer si un concept issu de domaines relativement éloignés de la sphère bibliothéconomique peut néanmoins être utile pour un travail sur ces fonds. Si c'est le cas, il faudra déterminer quelles caractéristiques de ces ensembles entretiennent des affinités avec les traits distinctifs des volumes de données massives que l'on intitule « mégadonnées ». Par ailleurs, et surtout pourrait-on dire, il faudrait préciser quelles méthodes de traitements élaborées par les informaticiens et les scientifiques présentent un intérêt pour l'exploitation et l'accès aux collections numérisées. Le parallèle entre les « mégadonnées » et les collections numériques n'est pas à envisager simplement comme une pure spéculation récréative, visant à nouer des liens artificiels entre une notion en vogue et les collections numérisées. On n'aurait pas fait grand chose, si l'on se contentait d'affirmer que les collections numérisées relèvent du « big data ». Il s'agit bien plutôt de penser quelles sont les méthodes et les outils dont le bibliothécaire pourrait s'inspirer pour permettre de nouveaux modes d'exploitation des collections numériques à large échelle, et déterminer si certaines des techniques liées aux « mégadonnées » peuvent s'appliquer avec profit aux fonds numérisées pour créer de nouvelles formes d'instrument de recherche.

L'approche retenue est donc essentiellement pragmatique, et quand nous disons interroger les relations qu'entretiennent big data et bibliothèques, il faut comprendre qu'il s'agit essentiellement de s'intéresser à l'apport potentiel des

[12]Emmanuelle Bermes, « Big data et bibliothèques », *Figoblog*, 13 janvier 2015. (Consulté le 3 août 2015) (disponible en ligne sur : <http://figoblog.org/2015/01/13/big-data-et-bibliotheques/>).

[13]Sur la bibliothèque de Singapour, voir Tao Ai Lei, « Singapore library mines big data », *Business information ASEAN*, septembre 2014, p. 2 à 5 (disponible en ligne sur : <http://bigdataanalytics.my/downloads/BI_Asean_Sept_2014.pdf>) (consulté le 4 août 2015).

techniques en usage à l'échelle des données massives pour l'accès, le traitement et la valorisation des collections numériques. Par « techniques en usage à l'échelle des données massives », qu'entend-on précisément dans ce travail ? On pense ici d'abord à la fouille de texte (text mining)[14], c'est à dire un ensemble de pratiques informatiques que l'on peut répartir en quatre types de tâches : la recherche d'information, la classification automatique de documents, l'annotation et l'extraction d'information[15]. Toutes ces opérations visent cependant le même objectif, extraire de l'information à partir d'ensembles conséquents de textes numériques. Comme le statut juridique du text mining fait l'objet de nombreux débats, au moment de la rédaction de ce mémoire, il sera nécessairement examiné plus en détail, mais notons par ailleurs que d'autres traditions méthodologiques d'analyses statistiques de grandes masses de texte peuvent également être appliquées aux collections numériques des bibliothèques : l'école française de statistique textuelle, également appelée textométrie, sera également évoquée. Cette approche centrée sur les techniques n'est pas d'ailleurs techniciste : si certains algorithmes d'analyse statistique peuvent indéniablement approfondir la connaissance que nous avons des collections, ils doivent être utilisés en ayant conscience de leurs limites, ce qui ne peut se faire qu'en ayant une compréhension minimale des logiques mises en œuvres.

Ayant ainsi défini l'optique et la portée de ce travail, comme exploration centrée sur les techniques du big data appliquées aux collections des bibliothèques, il nous a semblé qu'il pouvait se poursuivre principalement dans trois directions :

1) Tout d'abord, l'essor chez les chercheurs de courants de recherche investis dans l'analyse de très grandes masses de données, dont des données textuelles – ce que d'aucuns ont appelé les big digital humanities – implique le développement chez les bibliothécaires de services *ad hoc* pour répondre à une demande d'une partie de leur public.
2) Ensuite, ces mêmes techniques appliquées par les humanistes sur les collections de textes qu'ils fouillent peuvent être utilisées par des bibliothécaires pour offrir des descriptions des collections numériques.
3) Enfin ces techniques peuvent également être utilisées pour améliorer le traitement documentaire. On pense ici à des techniques telles que la classification automatique de textes ou à la reconnaissance d'entités nommées.

Les techniques liées au big data - comme le text mining - ont depuis longtemps attiré l'attention des bibliothécaires, et particulièrement dans le monde anglo-saxon, d'où nous sont parvenues quelques publications révélatrices sur les réflexions en cours. Déjà Eric Lease Morgan en 2012[16] s'interrogeait sur les possibilités qu'offrait la fouille de texte pour améliorer l'accès aux collections numérisées. Cet auteur appelait de ses vœux l'implémentation dans les

[14] Historiquement, la fouille de texte précède de plus d'une décennies l'émergence du thème des données massives. Il reste que l'augmentation des données produites et conservées a donné à cette technique une toute nouvelle importance, que l'on pense à ses applications pour analyser les millions de messages produits par Twitter par exemple. Il n'est donc pas exagéré de dire que, précédant le big data, la fouille de texte en est pourtant devenue l'une des techniques emblématiques, bien que certains auteurs comme Pierre Delort considèrent que seule l'induction de motifs à valeurs prédictives à partir de signaux faibles dans les données peut être considérée comme technique propre aux mégadonnées. infra p. 16.

[15] Cette description de la fouille de texte en quatre tâches est issue d'Isabelle Le Tellier, *Introduction à la fouille de texte*, Paris, Université de Paris 3 – Sorbonne Nouvelle, (non paginé) (disponible en ligne sur : <http://www.lattice.cnrs.fr/sites/itellier/poly_fouille_textes/fouille-textes.pdf>) (consulté le 20 novembre 2015).

[16] Eric Lease Morgan, « Use and understand : the inclusion of services against texts in library catalogs and « discovery systems ». *Library Hi Tech*, Vol. 30, Issue 1, 2012, p. 35-39.

bibliothèques numériques d'outils en provenance du traitement automatique du langage et du text mining. Pour lui, ces outils permettraient de faciliter l'usage et la compréhension (*use and understand*) des collections en plein texte numérique, en permettant, par exemple, de mesurer la taille d'un texte en nombre de mots, d'évaluer l'importance d'un concept à l'intérieur de ce texte, et par conséquent de comparer ce texte à d'autres du même corpus. Cet auteur associait l'usage des techniques de fouille de texte en bibliothèques à un véritable changement de paradigme, celui du passage d'une bibliothéconomie centrée sur la recherche de l'information à une bibliothéconomie facilitant la compréhension de textes dont on suppose désormais facilitée la découverte grâce aux moteurs de recherche – avec peut-être beaucoup d'optimisme. Telles seraient les promesses que des techniques associées aux données massives comme le text mining laisseraient espérer.

Il reste à signaler qu'eus égards à la complexité du sujet, il n'a pas paru inutile de consacrer une partie entière à la question de la définition des données massives. Car ce que les digital humanists appellent des approches big data ne s'identifie qu'imparfaitement à ce que des spécialistes en informatique qualifieraient ainsi, de même que le big data du marketing n'est pas celui de la recherche scientifique. Plus haut, nous avons dit que l'approche retenue était d'étudier l'usage des techniques associées aux données massives sur les collections numérisées des bibliothèques, et ce au détriment d'une autre approche qui aurait consisté à préciser immédiatement en quoi les collections numériques des bibliothèques relèvent des mégadonnées : si nous avons choisi de procéder ainsi, c'est que les collections des bibliothèques ne partagent pas toutes les caractéristiques des mégadonnées, et c'est également un point qu'il faudra justifier. D'autre part, il n'est pas certain que les données dont l'agrégation constitue le big data correspondent aux anciennes définitions de ce terme de données, ce qui, également, doit être expliqué pour justifier que les collections en pleins textes des bibliothèques puissent être qualifiées de données. Ce travail étant consacré aux techniques que l'on associe au big data - bien que la fouille de texte elle-même ait préexisté à la diffusion de ce terme -, il semblait nécessaire d'opérer un travail de clarification de ces différentes notions avant d'entrer dans le vif du sujet.

MEGADONNEES ET COLLECTIONS NUMERIQUES DES BIBLIOTHEQUES

Au regard des bibliothèques, le big data apparaît comme une notion issue du monde du commerce en ligne, et dont la pertinence pour réfléchir aux problèmes

soulevés par la gestion des collections numériques ne va pas de soi. Il est vrai que l'usage courant de l'expression identifie le big data au simple processus de recommandations algorithmique ou à l'exploitation commerciale des données personnelles[17], ce qui ne rend pas tout à fait justice à la complexité des débats autour des mégadonnées, tels qu'ils se développent dans les champs de la recherche scientifique et des sciences de l'information. Cette incompréhension et ces définitions réductrices sont facilitées par la polysémie d'une expression dont les significations varient selon les contextes et les locuteurs. C'est pourquoi, avant de pouvoir exposer en quoi les enjeux liés aux mégadonnées jettent un éclairage nouveau sur les collections électroniques, une entreprise préalable de clarification de la notion s'impose, qui permettra de travailler sur la base de définitions rigoureuses, plutôt que de faire fond sur le vague d'idées préconçues au sujet de ce qu'est ou n'est pas le big data.

Au terme de cette première étape, qui a pour but de produire des définitions rigoureuses des mégadonnées, il sera alors possible de relever quelles caractéristiques, appartenant à quelle définition semblent caractériser adéquatement la situation des collections numériques, permettent de relier le phénomène du big data aux problèmes des collections numériques, et donc de comprendre l'intérêt de faire usage de techniques spécifiques aux données massives sur les collections numériques.

À LA RECHERCHE D'UNE DEFINITION POUR LES BIG DATA

Fournir une définition liminaire des mégadonnées[18] n'est pas chose aisée, car pour répandue qu'elle soit, cette notion attend toujours de recevoir une définition à la fois rigoureuse et universellement admise. Plutôt que d'un concept aux contours et à l'extension clairement définis, on parlera à son égard plus volontiers d'un ensemble notionnel, où, par une équivoque significative, le terme de mégadonnées peut être à la fois appliqué à un objet – qui serait le volume de données dont la grandeur justifierait cette appellation-, ou aux méthodes d'analyse spécifiques qui leur sont appliquées pour en tirer une information exploitable. Significativement, une synthèse récente de la littérature scientifique sur le sujet[19] ne distinguait pas moins de trois grandes catégories de définition pour les mégadonnées, et trouvait bon d'en rajouter une de son propre cru.

Nous avons choisi de retenir cette dernière typologie comme fil conducteur de l'exposé. Par conséquent, rappelons ici brièvement les différentes orientations définitoires qui y sont recensées pour caractériser le big data : une première perspective retient comme détermination principale du phénomène un changement objectif et quantitatif dans la taille des données. C'est une définition centrée sur l'objet des big data, qui conduit les auteurs de la synthèse à la qualifier de perspective centrée sur le produit (*product-oriented perspective).* À cette première orientation définitoire s'oppose une autre qui elle retient comme trait spécifique du big data une évolution significative de la capacité à analyser ces données, et à en

[17]On en voit un exemple dans cette brève consacrée à la politique big data de la SNCF : Christophe Auffray, « Big Data - Voyages-sncf.com se transforme grâce aux données », *ZdNet*, 25 novembre 2015 (disponible en ligne sur : <http://www.zdnet.fr/actualites/big-data-voyages-sncfcom-se-transforme-grace-aux-donnees-39828706.htm>) (consulté le 27 novembre 2015).

[18]Les termes de « big data », de « mégadonnées » et de « données massives » seront employées indifféremment dans ce travail.

[19]Hamid Ekbi, Michael Mattioli, Inna Kouper, *et alii*, « Big data, bigger dilemnas : a critical review. », *Journal of the Association for Information Science and Technology,* Août 2015, p. 1523-1545.

tirer parti. Dans la mesure où il est ici question des processus informatiques d'analyse mis en œuvre pour analyser les données, cette définition est centrée sur les processus (*process-oriented perspective*), et met l'accent sur les innovations qui permettent de penser le big data comme une percée technologique dans l'analyse des données, qui autoriserait désormais de traiter « l'opacité, le bruit, et les relations entre ces données », alors même que leur masse ou leurs caractéristiques intrinsèques (comme l'absence de structuration) défiaient précédemment l'analyse. Cette définition retient donc comme critère définitoire une amélioration des techniques, pour mieux gérer l'analyse, le traitement, et le stockage de ces données. Enfin, le dernier type de définition mis en avant par les rédacteurs de cet article est centré sur le problème cognitif que pose le développement des mégadonnées aux chercheurs : les mégadonnées sont dès lors définies non pas par leur objet, ni par les processus impliqués, mais par l'excès qu'elles représentent par rapport au capacité d'analyse des êtres humains. Une dernière orientation, celle suggérée par les auteurs mêmes de l'article, définit le big data à partir de ses répercussions sociales. Toutefois, nous ne nous y attarderons pas, comme en dehors de notre problématique.

A partir des trois catégories esquissées par les auteurs de cet article, reprenons le fil de notre exposé. On verra que les différentes orientations proposées ne sont pas forcément exclusives les unes des autres : la distinction a surtout une valeur heuristique, en permettant de mettre en avant des orientations prioritaires retenues par les acteurs. Il va de soi par ailleurs que des personnes caractérisant le big data par une méthode peuvent également évoquer des caractéristiques objectives des données, et vice-versa.

Les mégadonnées comme changement dans les caractéristiques objectives des données

Cette première façon de caractériser le phénomène a trouvé sa formulation canonique dans une définition répandue[20], qui s'appuie sur une trilogie de termes partageant la même initiale : selon cette règle, dite des « 3V », le big data serait une évolution dans le rythme de croissance des données, qui en modifierait à la fois le volume, la vélocité, et la variété. Cette définition, bien que contestée[21], a le mérite de dépasser une simple caractérisation du « big data » par la taille des données concernées, bien que cette dernière soit toujours la première caractéristique citée. Développons brièvement les éléments de cette définition, qui se concentre sur ce que le big data signifie pour les données.

Le volume rappelle que l'invention du concept cherche d'abord à cerner une évolution significative dans notre capacité à capter, créer et stocker les données. Il est d'usage de citer à ce sujet certains chiffres particulièrement impressionnants : on aime ainsi à rappeler que chaque jour, le réseau social Facebook rassemble environ 500 teraoctets de données, ou encore que Google analyse environ 24 petaoctets de données[22]. Dans un autre domaine, celui de la recherche scientifique, la taille des données captées connaît elle aussi une inflation spectaculaire : le

[20]Ce modèle figure notamment en bonne place des pages anglaises et françaises que Wikipedia consacre au « big data ». Voir Article « Big data », *Wikipedia* (disponible en ligne sur : <https://fr.wikipedia.org/wiki/Big_data>)(<consulté le 16 novembre 2015).

[21]Voir Pierre Delort, *Le big data*, Paris : Presse universitaire de France, 2015, p. 5.

[22]Pour ces chiffres, voir Hamid Ekbi, Michael Mattioli, Inna Kouper, *et alii, Ibidem,* Août 2015, p. 1526.

radiotéléscope « Square Kilometre Array » devrait ainsi produire 50 teraoctets de données analysées par jour lors de sa mise en service[23].

Pourtant, la masse exprimée n'est pas tout. Ce volume, qui paraît en premier lieu un trait définitoire essentiel, ne rend pourtant que des services bien décevants pour aider à préciser ce que l'on entend par données massives : cette grandeur, notion relative entre toutes, appelée à subir une prompte dévaluation en raison de l'extension indéfinie des capacités de stockage, n'a d'ailleurs que rarement fait l'objet de tentative de quantification précise.

La vélocité fait, elle, référence à la vitesse de production de ces données : dans le cas de Google, les données récoltées ne sont rien moins que l'ensemble des requêtes lancées à travers le moteur de recherche, créées à la vitesse où des milliers d'internautes tapent leurs requêtes dans la fameuse barre du service ; dans le cas d'un site marchand, les données sont créés à la vitesse de la navigation de l'usager sur le site.

Enfin, la Variété : cette dernière notion, particulièrement intéressante, peut être comprise de façons diverses. On peut, d'une part, la comprendre comme une évolution des caractéristiques formelles de ces atomes dont l'agrégation forme le big data, les donnée elle-même. Auparavant, la donnée se signalait par des caractéristiques précises : elle était, selon une définition parmi d'autre, « toute information structurée de manière reconnaissable »[24]. Pour prendre un exemple, les usagers de bases de données savent bien que chaque unité d'une table est composée de plusieurs rubriques semblables : les données des bases de données, comme les données évoquées par la définition ci-dessus, étaient des ensembles d'information structurées dès leur création pour présenter une homologie de forme dans un ensemble sériel. Cependant, les données du big data ne répondent plus que partiellement à cet impératif d'homogénéité et de structuration préalable ; quand la donnée peut indifféremment être le message d'un réseau social, une vidéo postée sur une plate-forme communautaire, ou encore les requêtes lancées sur un moteur de recherche, il semble évident que la structuration n'est plus la caractéristique principale d'une donnée. La Variété est donc la caractéristique d'un âge informatique où tout est donnée, contenu structuré ou non-structuré, et cette caractéristique intéresse tout particulièrement les bibliothèques, pour qui la donnée est d'abord le texte, c'est à dire un type de données pour qui la structuration intégrale par l'usage d'un marquage sémantique comme XML/TEI reste exceptionnelle. On peut qualifier la masse des textes numériques au mieux de semi-structurée, par l'usage de mises en forme différentes selon les portions de texte, et la répartition du texte en paragraphes distincts.

Toutefois, cette variété peut être entendue dans un autre sens, celui de la mise en relation de jeux de données hétérogènes afin de repérer des corrélations intéressantes. Nous nous rapprochons déjà ici de la perspective procédurale, qui sera évoquée par la suite. On en a vu précédemment un exemple, avec celui des pompiers de New-York. Mais on peut penser également à l'exemple du service de police de la ville de Santa Cruz[25] : leur programme d'analyse de données met en relation des jeux de données d'origine diverses (structure architecturale des différentes zones urbaines, historique des délits commis) pour mieux répartir les forces de police.

[23]Article « Big data », *Wikipedia* (disponible en ligne sur : <https://fr.wikipedia.org/wiki/Big_data>)(<consulté le 16 novembre 2015).

[24]Lynda Kellam et Katharin Peter, *Numeric data services and sources for the general reference librarian*, Oxford, Chandos Publishing, 2011. p. 7-8. Cité dans Raphaëlle Lapôtre, *Ibidem*, p. 12.

[25]Gilles Babinet, *Big data, penser l'homme et le monde autrement*, Paris, le Passeur, 2015, p.42.

Tels sont les 3V de la fameuse règle des 3V. La commodité de cette définition trinitaire lui a assuré une large diffusion. Cette caractérisation n'épuise pourtant pas son objet, d'autant qu'on la tire d'un rapport du cabinet de conseil Meta Group dont la parution remonte à 2001, bien avant que le phénomène émergeant que l'on cherchait à appréhender ne soit baptisé du nom de big data, et que ses contours et conséquences soient mieux cernés. On complétera donc ce premier aperçu par deux définitions complémentaires. L'une, d'allure pragmatique, aborde la question des données massives par une approche procédurale sous l'angle de leurs conséquences pour la gestion et l'analyse : les mégadonnées consisteraient alors en ensembles de données dont la nature et la structuration sont telles qu'elles ne peuvent être gérées par des moyens conventionnels, c'est à dire « des outils de gestion de bases de données ou de gestion de l'information »[26]. Envisagé sous cette angle, serait mégadonnées un volume de données dont la nature et la structuration prennent au dépourvu les outils dit traditionnels de gestion, comme les bases de données relationnelles, où l'information est structurée sous la forme de tables et de colonnes. Cette caractérisation des mégadonnées par leur mode d'administration est précieuse, car elle permet de mieux saisir le caractère labile d'une notion qui tend à embrasser tout ensemble de données qui, pour un acteur, dépasse ses capacités de gestion.

Cependant, le terme big data, pour certains auteurs, ne se contente pas de définir un objet, mais désigne surtout une méthode.

Les mégadonnées comme nouvelles méthodes d'analyse des données : l'approche procédurale

Nous avons vu plus haut qu'une autre approche caractérisait le big data moins par les caractéristiques objectives des données que par l'application à ces données de nouvelles méthodes d'analyse, c'est à dire par de nouvelles façons de les faire parler. La caractéristique première du big data serait, dès lors, d'être une méthode permettant le repérage de « signaux faibles » à l'intérieur de données massives, c'est-à-dire de phénomènes de corrélations uniquement perceptibles à l'échelle des grandes masses de données. Cette approche est celle retenue par exemple par Pierre Delort, qui a proposé la définition suivante du big data :

> Le Big Data consiste à créer en exploratoire et par induction sur des masses de données à faible densité en information des modèles à capacité prédictive[27].

Tirer un enseignement exploitable d'une masse de données partiellement informes – c'est à dire dont le contenu n'a pas été structuré au préalable pour répondre à une demande, et où, par conséquent les informations de valeurs sont présentes de façon clairsemées -, tel serait le big data, appréhendé dans une perspective qui prend en compte également méthode d'analyse et qualification objective de la masse de données. Cette définition se comprendra mieux en revenant à l'exemple dont se sert Pierre Delort pour l'illustrer sa proposition, celui du programme Google Flu, élaboré par l'entreprise californienne en collaboration avec un organe du ministère de la santé américain, le Center for desease control

[26]Article « Big data », *Wikipedia* (disponible en ligne sur : <https://fr.wikipedia.org/wiki/Big_data>)(<consulté le 16 novembre 2015).

[27]Pierre Delort, *Le big data*, Paris, Presse universitaire de France, 2015, p. 42.

(CDC), pour permettre d'évaluer quasiment en temps réel la propagation des épidémies de grippe sur le territoire américain. Ce programme, présenté dans un article publié dans la revue *Nature*[28], s'est appuyé sur la mise en relation des cinquante millions de requêtes les plus fréquentes lancées par les Internautes dans le moteur de recherche avec les données remontées par les médecins américains signalant au ministère les possibles cas de grippe. Au terme de cette analyse, les chercheurs ont présenté une liste de 45 mots clefs dont la recherche avait révélé que leur utilisation via le moteur de recherche était corrélée avec une avancée de l'épidémie grippale dans la zone géographique d'où étaient lancées les requêtes. Les chercheurs ont donc présenté un modèle d'analyse permettant, en temps réel et dans une certaine mesure, de suivre l'avancée d'une épidémie grippale sur le terrain, et ce à partir des requêtes lancées par les Internautes dans le moteur de recherche.

Les caractéristiques du Big data, quand on le présente comme une nouvelle méthode d'analyse, sont toutes rassemblées dans ce cas : un ensemble de données hétérogènes, non structurées a priori pour l'usage qui en sera finalement fait – ici rien de moins que des milliards de logs de recherches passées sur le moteur, qui n'étaient pas destinées à être utilisées pour détecter les avancées d'une épidémie de grippe -, une faible densité informationnelle de ces données non structurées – seules une partie des requêtes pouvaient être significatives -, et le croisement entre elles de données hétérogènes – les informations en provenance du CDC et les logs du moteur de recherche. Enfin, à partir de cette ensemble, un raisonnement par induction permet d'établir des règles prédictives.

La plus importante de ces caractéristiques est certainement la faible densité informationnelle, conséquence naturelle de l'absence de structuration : de fait, il n'est pas rare que le big data soit présenté d'abord comme l'art de repérer des signaux faibles dans une masse de données.

Les mégadonnées comme problème cognitif : un excès de donnés par rapport aux capacités d'analyse des chercheurs

Cette dernière remarque permet de comprendre des usages du terme de big data, qui échappent aux deux formulations précédentes, mais expliquent très bien certains points que nous verrons dans la seconde partie de ce travail, consacrée aux usages du big data dans les sciences humaines et sociales. Le big data, comme nous venons de le montrer, peut à la fois être considéré comme un type particulier de données, ou encore comme un type de méthode. Mais il peut, de façon plus courante, désigner une masse de données telle qu'elle excède les facultés d'analyse des chercheurs. Et à cette aune, d'éminents spécialistes d'histoires littéraires peuvent qualifier d'analyse à l'échelle des mégadonnées des recherches qui concernent des corpus gros de seulement quelques milliers de textes littéraires[29]. A ce sujet, on peut citer S. Graham, I. Milligan et S. Weingart[30] :

> Le big data pour des chercheurs en littérature, cela peut être une centaine de romans (« la masse sans lecteurs »), pour des historiens, cela

[28]Jeremy Ginsberg, Matthew H. Mohebbi, *et alii*, « Detecting influenza epidemics using search engine query data », *Nature*, n°457, p. 1012-1014, 19 février 2009 (disponible en ligne : http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html) (consulté le 13 novembre 2015).

[29]Voir par exemple Matthews L. Jocker, *Macroanalysis*, Springfield, University of Illinois Presse, p. 20.

[30]S. Graham, I. Milligan et S. Weingart. *The historians's macroscope : big digital history,* p. 16 (Disponible en ligne : <http://www.themacroscope.org/?page_id=597>) (consulté le 16 décembre 2015).

> peut-être un tableau entier de listes d'expéditions (…) Pour nous, chercheurs en sciences humaines, le big du big data est défini par l'œil du chercheur. Si vous disposez de plus de données que vous ne pourriez en lire par vous-même dans un laps de temps raisonnable, ou si ces données nécessitent une analyse informatique pour que vous puissiez les interpréter, alors les données sont déjà massives ![31]

Il faut donc se rappeler qu'à côté des définitions techniques, élaborées par les informaticiens, voisine cette perception intuitive du big data, que l'on retrouve souvent sous la plume des spécialistes des humanités numériques : également valide, elle se fonde pragmatiquement sur la difficulté à analyser certaines données, sans en qualifier le seuil ni la forme. Le big data, dès lors, est simplement le nom assignée au seuil à partir duquel les méthodes traditionnelles de recherche, basées sur une lecture attentive des textes, ne peuvent plus être mises en œuvre.

Conclusion provisoire : le big data et ses conséquences sur définition des données

En rassemblant ces informations puisées à des sources différentes, nous avons désormais une idée plus précise des mégadonnées, qu'il s'agira de confronter avec les masses documentaires traitées en bibliothèque pour établir si ce concept allogène est pertinent en bibliothéconomie. Les mégadonnées représentent donc un ensemble de données dont le volume et la nature empêchent le stockage, le traitement et l'analyse par des moyens conventionnels, tels que les bases de données relationnelles ;parmi les caractéristiques principales qui expliquent cet état de fait, on signalera surtout leur absence de structuration, leur volatilité, et leur hétérogénéité. Mais le big data est également une méthode, celle qui consiste, par l'usage de moyens informatiques importants, à induire des modèles prédictifs de masses d'informations non structurées, et à partir de signaux faibles dispersés dans cette masse, qui ne seraient pas perceptibles, si le volume de données n'était pas si important[32]. Par ailleurs, on peut d'ailleurs signaler qu'une acception plus triviale du terme a également cours, qui mesure les mégadonnées à l'aune de l'impuissance d'un public déterminé à les interpréter sans recourir à l'assistance d'une analyse informatisée.

On verra plus bas ce qui, dans les caractéristiques des collections électroniques des bibliothèques, correspond à ces traits définitoires des mégadonnées. Pour l'instant, on peut se demander quelles conséquences le big data peut avoir sur la définition même de la donnée, l'unité de base dont l'accumulation finit par former les mégadonnées, et ce point est également important pour

[31] « How big is big?" we rhetorically ask: big data for literature scholars might mean a hundred novels ("the great unread"), for historians it might mean an entire array of 19th century shipping rosters, and for archaeologists it might mean every bit of data generated by several seasons of field survey and several seasons of excavation and study – the materials that *don't* go into the Geographic Information System. For us, as humanists, *big is in the eye of the beholder*. If it's more data that you could conceivably read yourself in a reasonable amount of time, or that requires computational intervention to make new sense of it, it's big enough! » S. Graham, I. Milligan et S. Weingart. *Ibidem*, p. 16

[32] Plus les données sont informes et à faible densité informationnelle (en tout cas par rapport au centre d'intérêt de l'observateur), plus leur volume doit devenir important pour vérifier une hypothèse concernant des corrélations entre données. C'est bien le paramètre du volume qui va permettre de percevoir des corrélations infimes entre phénomènes, qu'un nombre trop faible de données ne permettrait pas de distinguer. Ce lien entre les « signaux faibles » et l'importance du volume est explicitement relevé par Pierre Delort, *Ibidem,* p.46 : « Pour ces données clairsemées qui sont caractéristiques du Big data, il a été empiriquement montré que le volume améliore les modèles prédictifs »

déterminer si les collections numériques des bibliothèques entrent, ou pas, dans le champ de la définition.

Cette donnée dont l'accumulation forme le big data ne peut plus se limiter aux définitions, excellentes, qu'on peut en donner par ailleurs dans d'autres contextes. Par exemple, insistant sur les différences entre données, informations et connaissances, Serge Abiteboul définissait ainsi cette notion :

> Une donnée est une description élémentaire, typiquement numérique pour nous, d'une réalité. C'est par exemple une observation ou une mesure.[33]

Pour cette définition, on voit que la donnée par excellence est la valeur d'une variable statistique. On citera également cette autre : « toute information structurée de manière reconnaissable »[34], intéressante car elle met en avant la notion de structure récurrente. Le big data n'a pas rendu caduques ces deux définitions, qui peuvent toujours s'appliquer à certaines des données qui en composent les agrégats. Ainsi les données personnelles liées aux historiques d'achat sont bien des données structurées : chaque information correspond à un champ dans une base de données. Il reste que ces deux définitions ne s'appliquent plus qu'à une partie de la matière qu'analysent les data scientist, qui ont annexé à leur domaine une masse de données dont la caractéristique majeure est d'être pas, ou peu structurée : les données textuelles. Ces données textuelles, qui intéressent au premier chef les bibliothèques, sont en effet l'un des premiers terrains d'investigation pour les data scientist[35], qu'il s'agisse d'analyser les sentiments exprimés sur les messages des réseaux sociaux pour veiller sur la réputation d'une entreprise, ou encore d'analyser automatiquement les messages de réclamations qui parviennent à un service clientèle.

Pour résumer l'impact du big data sur la définition de la donnée elle-même, on dira que le terme recoupe désormais tout document numérique susceptible d'être traité, interprété et analysé par un programme informatique. C'est cette dernière caractéristique qui définit désormais la donnée, plutôt que le fait d'être structurée de manière reconnaissable. Dans cette perspective, tout objet numérique est susceptible de devenir donnée « par destination », et notamment le texte des collections électroniques, qui, pour être analysé, pourra soit subir un ensemble de traitement visant à le convertir sous forme numérique – et donc à dissocier les éléments du texte pour les réduire à des variables statistiques -, soit se voir appliquer des techniques de traitement du langage naturel, visant à l'extraction d'élément du texte par reconnaissance sémantique. Mais dans les deux cas, le texte même est devenu une donnée.

C'est pourquoi dans ce travail sur le big data et les collections des bibliothèques, les données dont nous traiterons seront à la fois les métadonnées, et le texte même des collections, dans la mesure où le développement des techniques

[33]Serge Abiteboul. *Sciences des données : de la logique du premier ordre à la toile : Leçon inaugurale prononcée le jeudi 8 mars 2012* (disponible en ligne sur : <http://books.openedition.org/cdf/529#bodyftn5>) (consulté le 5 novembre 2015).

[34]Lynda Kellam et Katharin Peter, *Numeric data services and sources for the general reference librarian*, Oxford, Chandos Publishing, 2011. p. 7-8. Cité dans Raphaëlle Lapôtre, *Ibidem*, p. 12.

[35]Sans avoir reçu pour le moment de définition bien établie, la désignation de data scientist fait référence aux experts de la data science, qui elle est définie comme « un champ de recherche interdisciplinaire concernant les processus et les systèmes pour extraire de la connaissance ou des aperçus à partir de données présentées sous des formes diverses, structurées ou non structurées. La data science est la continuation de certains champs de l'analyse de données comme les statistiques, l'exploration de données et l'analyse prédictive ». « Data science », *Wikipedia* (en ligne : <https://en.wikipedia.org/wiki/Data_science>). Les data scientist s'intéressent donc aux problématiques des données massives sous leur aspect de nouvelle méthode d'analyse des données.

d'analyse textuelle liées au big data ont fait de ces textes des données comme les autres. Cependant, outre le fait que le texte, qui est la matière première des collections hébergées par les bibliothèques, puisse désormais être considéré comme une donnée, on peut se demander si d'autres caractéristiques des collections numériques peuvent justifier une comparaison entre les problématiques du big data et celles des collections numériques.

COLLECTIONS NUMERIQUES DES BIBLIOTHEQUES ET MEGADONNEES : QUELS RECOUPEMENTS ?

Analogie des caractéristiques

La revue des différentes définitions du big data à laquelle nous venons de procéder nous a, semble t-il, emmené bien loin des questions qui intéressent la bibliothéconomie. Il nous appartient donc désormais de revenir sur ce terrain, pour examiner quelles relations peuvent exister entre les ensembles caractérisés de mégadonnées et les collections numériques. Nous venons déjà de constater, en voyant comment le big data affectait la notion de données, qu'il fallait compter au nombre des données hébergées par les bibliothèques le corps des textes des collections numériques, en plus des métadonnées et des chiffres sur l'activité de la bibliothèque, qui viennent immédiatement à l'esprit. Nous allons voir désormais que les collections des bibliothèques présentent d'autres analogies objectives avec les mégadonnées telles que nous les avons définies. Ces collections sont en effet marquées par une absence relative de structuration – puisqu'il s'agit de textes le plus souvent semi-structurés, faute d'un usage systématique d'une grammaire d'encodage de type XML/TEI -, par l'hétérogénéité de formats, et, enfin, par une faible densité informationnelle, qui découle naturellement du fait qu'il s'agisse de texte, un vecteur de communication où les structures grammaticales, les éléments de syntaxes, éventuellement la rhétorique, accompagnent l'information, qui ne se présente pas seule comme dans une base de données. Une fois décrite et justifiée cette analogie, il sera temps d'examiner comment les techniques du big data peuvent s'appliquer aux collections.

Les collections électroniques des bibliothèques sont classiquement réparties[36] en quatre ensembles : les abonnements de périodique électronique de niveau recherche, les collections numériques patrimoniales, les archives de site Internet rassemblées au titre du dépôt légal du Web, et enfin l'offre culturelle numérique d'e-books, de revues, de titres de presse, de VOD. Nous pouvons pour commencer étudier ces ensembles au prisme de chacun des attributs qui composent la règle des « 3V » des mégadonnées.

Intéressons nous tout d'abord au volume que représentent en bibliothèques les collections électroniques. Tout en reconnaissant le caractère insatisfaisant de cette seule approche, il serait paradoxal de l'évacuer d'emblée dans un travail consacré à la question des données massives. Sans surprises, les collections des bibliothèques électroniques ne peuvent prétendre être comparées aux gigantesques entrepôts de données de la recherche de pointe, ou même à ceux que gèrent les grands acteurs d'Internet. Pour s'en tenir aux collections patrimoniales, une bibliothèque numérique d'une taille importante comme l'américaine Hathi Trust,

[36] Voir Emmanuelle Bermès, Frédéric Martin. Le concept de collection numérique. *Bulletin des bibliothèques de France*, n° 3, 2010. (disponible en ligne :<http://bbf.enssib.fr/consulter/bbf-2010-03-0013-002>)

représentait en 2012 environ 450 Teraoctets pour dix millions de livres numérisés[37], à comparer avec les taux d'accroissement quotidiens d'acteurs comme Twitter (7 To par jours[38]). Une bibliothèque numérique « de niche » comme celle de la BIUS, Medic@, consacrée aux collections patrimoniales d'histoire de la médecine, rassemble 4,6 To pour 15000 volumes[39]. Bref, au risque de rappeler une vérité d'évidence : si les bibliothèques sont confrontées avec des problématiques similaires à celle des données massives, ce n'est pas au niveau du pur et simple volume informatique des collections, puisque ces établissements, en tant que dépôts de données, ont depuis longtemps été débordés par la fécondité informationnelle du monde numérique. La langue des journalistes a d'ailleurs entériné ce déclassement des bibliothèques dans la hiérarchie des conservatoires de données, en faisant de la bibliothèque du Congrès une unité de mesures informatiques parmi d'autres, correspondant à environ 10 teraoctets de textes non compressés[40].

Notons cependant qu'il n'en va pas tout à fait de même pour toutes les régions des collections électroniques : ainsi l'archivage d'Internet, fût-il envisagé uniquement sur le seul plan du volume informatique, est bien de plain-pied avec le monde des mégadonnées, puisque son objet est précisément d'archiver l'un des domaines constitutifs du big data. Chaque année, la mission de collecte au titre du dépôt légal du Web impose à la BNF de stocker quelques 100 teraoctets supplémentaires[41]. Encore cet accroissement doit-il s'apprécier en gardant à l'esprit que la BNF n'archive que les sites appartenant au nom de domain .fr. Un entrepôt généraliste tel que l'Internet Archive déclare archiver ainsi entre 50 et 60 teraoctets par semaine[42], et conservait le premier juillet 2015 23 petaoctets. On voit par là que même à travers cette question triviale du volume informatique, les bibliothèques sont bien concernées par cet aspect du big data.

Mais nous avions noté plus haut qu'au-delà du volume informatique, ce qui faisait le big data, ce qui impliquait ses problèmes et ses techniques particulières, ce n'était pas simplement le poids des collections traitées, mais plus précisément certaines caractéristiques de la structure des données concernées, qui tout à la fois les rendent difficiles à gérer avec des moyens dits conventionnels et nécessitent pour leurs exploitations des moyens *ad hoc*. Il semble que ce soit en ce point que les caractéristiques du big data et celles des collections électroniques se rejoignent, du fait que ces deux ensembles partagent deux spécificités, qui sont la variété et la faible densité informationnelle. Il n'est peut-être pas nécessaire de revenir longuement sur la variété : contentons nous ici de rappeler que dans le contexte des bibliothèques, il s'agit ici d'évoquer ici des données de type non structurées, non formatées en fonction de leur utilisation ultérieure, dissemblables des données préformatées des bases de données. Les collections électroniques des bibliothèques

[37]Voir la présentation suivante : Tom Burton-West, *HathiTrust Large Scale Search : scalability meets Usability*, diapositive 2 (disponible en ligne sur : <https://www.hathitrust.org/papers_and_presentations>) (consulté le 5 novembre 2015).

[38] Voir *CNRS International magazine*, n° 28, janvier 2013, p. 21 (disponible en ligne : <http://www.cnrs.fr/fr/pdf/cim/CIM28.pdf>, consulté le 6 décembre 2015).

[39]Entretien avec Claire Ménard, chargé de ressources numériques à la BIUS.

[40]Voir Leslie Johnston, « How many libraries of congress does it take », 23 mars 2012, (disponible en ligne : https://blogs.loc.gov/digitalpreservation/2012/03/how-many-libraries-of-congress-does-it-take/) (consulté le 2 novembre 2015).

Voir l'article que Wikipedia consacre à la question des unités de mesure atypique, disponible en ligne : <https://en.wikipedia.org/wiki/List_of_unusual_units_of_measurement> (consulté le 6 novembre 2015).

[41]Information donnée par M. Peter Stirling, chargé de collections numériques au service du dépôt légal du Web de la BNF.

[42]Voir la section « Questions fréquemment posées » du site de l'Internet Archive, disponible à cette adresse : <https://archive.org/about/faqs.php> (consulté le 6 décembre 2015).

étant composées de textes, soit de données au mieux semi-structurées (sauf si un marquage de type XML/TEI leur a été appliqué, ce qui reste l'exception), on peut les qualifier de données majoritairement non structurées ou semi-structurées. Par ailleurs, autre spécificité qui rejoint la problématique du big data, les textes des bibliothèques sont à faible densité informationnelle : cette caractéristique est le corollaire de la précédente, et il faut entendre par là que le caractère textuel des données conservées par les bibliothèques entraîne pour conséquence que l'information brute que le texte délivre s'y trouve qu'immergée dans un document dont une grande partie est constituée d'informations non pertinentes. Dans le cas d'un article scientifique, toutes les informations éparses dans le texte ne sont pas également pertinentes ; le résultat d'une analyse est, par exemple, précédé par la description du protocole de l'expérience. Le texte est, par nature, un type de donnée prolixe, où l'information tant désirée ne se trouve pas exposée à nue.

Hétérogénéité des formats -ici des formes documentaires -, absence complète ou partielle de structuration, présence d'informations pertinentes noyées dans la masse : telles étaient quelques-unes des caractéristiques signalées précédemment au-sujet des mégadonnées. La question des signaux faibles était particulièrement mise en avant par Pierre Delort, qui faisait de la faible densité informationnelle l'une des caractéristiques éminentes des mégadonnées.

Envisagé d'après les caractéristiques objectives des données hébergées par les bibliothèques, on voit qu'avec des nuances, la notion de mégadonnées avec ses attributs s'y applique bien. Cependant, c'est encore du point de vue des chercheurs, et des limites qui sont celles de leurs capacités d'analyse sans l'aide de moyens informatiques, que la comparaison est la plus pertinente. Nous en revenons donc ici à la question du big data comme défi cognitif posé par le déluge des données à la compréhension humaine. Et cette question se pose en particulier dans un secteur des collections numériques, qui est celui des abonnements de périodique électronique de niveau recherche. Dans ce secteur particulier des collections, l'accroissement mondial du volume de la littérature scientifique pose avec une acuité particulière la question de l'assistance informatisée de la maîtrise de ce volume[43].

L'importance des techniques informatiques d'analyse des collections

Ce phénomène d'accroissement de la littérature scientifique est bien l'origine de l'intérêt du monde de la recherche pour les techniques que l'on rassemble sous l'appellation de Text et data mining, abrégé en TDM par la suite. Un exemple suffira à exposer en quoi la question de la fouille de texte est devenue importante pour la question de l'accès aux collections. Dans la recherche scientifique, les données s'accumulent, dispersées à travers les articles publiées en nombre toujours croissant[44]. Ainsi, en ce qui concerne le domaine de la génomique

[43]Pour trivial et partagé que puisse paraître ce constat, quantifier cette augmentation du nombre d'articles publiés représente d'ailleurs une gageure, dans la mesure où les bases de données scientifiques réputées les plus exhaustives – comme le Thomson Web of Science - ne captent elles-mêmes que partiellement les résultats de la recherche mondiale. Une des dernières hypothèses avancées (Lutz Bornmann, Rüdiger Mutz. « Growth rates of modern science : a bibliometric analysis based on the number of publications and cited references. » (disponible en ligne sur : <http://arxiv.org/ftp/arxiv/papers/1402/1402.4578.pdf>) pour parvenir à une description chiffrée du phénomène suggérait, avec de nombreuses réserves, que le nombre d'articles publiés augmenterait de 8 à 9 % chaque année, ce qui induit par conséquent un doublement tous les neuf ans du nombre global d'articles scientifiques publiés dans le monde.

[44](Éditorial sans mention d'auteur), « Gold in the text ? », *Nature*, Vol. 483, 8 mars 2012 (disponible en ligne sur : <http://www.nature.com/nature/journal/v483/n7388/full/483124a.html>) (consulté le 23 décembre 2015).

chaque jour 2000 articles sont ajoutés à la base de données Medline, l'un des principaux outils de travail des chercheurs de cette discipline [45]. Ces données éparses, si elles étaient combinées et analysables dans leur ensemble, pourraient former un matériau fécond pour la recherche.

Ce problème d'abondance et de dispersion des données semble avoir été particulièrement sensible dans le domaine du décryptage du génome humain, puisque c'est de cette branche de la recherche que vient l'une des premières initiatives notables en matière de fouille de texte appliquée à la littérature scientifique. Une équipe constituée de chercheurs anglais et américains a cherché à y remédier, dans le cadre d'un projet intitulé « Text2genome » [46]. Selon les chercheurs impliqués dans cette recherche, ce domaine génère des articles où les données se trouvent en faible densité[47], problématique typique des mégadonnées comme nous l'avons vu. Pour y remédier, une équipe de scientifiques s'est attelée à élaborer un programme qui permette de lier des articles scientifiques aux gènes dont ils traitent. Cela implique qu'un algorithme de fouille de texte analyse les textes entiers de milliers de publications afin de repérer toutes les séquences d'ADN, les extraire, et enfin établir des liens entre l'article cité et le gène évoqué. La tâche de trouver des compte-rendus d'expérience pour une séquence d'ADN précise est ainsi grandement facilitée pour le spécialiste.

Le projet text2genom est également important car il s'agit de la première démarche de ce type à s'être heurtée aux restrictions importantes fixées par les éditeurs de littérature scientifique à l'exploitation par TDM de la littérature scientifique.

Mais outre la question de la maîtrise d'une masse documentaire en expansion, ce sont également de nouvelles méthodes de recherche qui expliquent l'intérêt que témoignent les scientifiques aux techniques de fouille de texte appliquées à la littérature de niveau recherche. Ces techniques peuvent ne pas seulement consister en procédés mis en œuvre pour accéder aux résultats de la recherche sous forme synthétique, mais également représenter des moyens de produire de nouveaux résultats, en permettant de déceler des liens de corrélations ou de causalités (entre maladies, entre substances médicinales, entre symptômes) dans la masse des articles répertoriant les résultats empiriques de la recherche. Dans la lignée des travaux de Don R. Swanson sur la découverte scientifique basée sur l'examen de la littérature[48], les chercheurs peuvent formuler des hypothèses de recherche à partir des collections d'articles scientifiques, c'est-à-dire, par l'examen minutieux des résultats de la recherche empirique tels qu'ils s'exposent dans les articles publiés, chercher à établir des liens de corrélations ou de causalités entre des concepts ou des entités entre lesquelles aucun lien direct n'a été établi au cours des expériences. Il s'agit, selon la formule de Swanson lui-même, de lier entre elles des connaissances disjointes car publiées dans des articles séparées : selon la

[45] Maximilian Haeussler, Martin Gerner, M. Casey Bergman, « Annotating genes and genomes with DNA sequences extracted from biomedical article », *Bioinformatics*, n° 27, 2011, p. 980-986 (disponible en ligne sur : <http://bioinformatics.oxfordjournals.org/content/27/7/980.full>) (consulté le 23 décembre 2015).

[46] On peut consulter une présentation de ce projet sur son site officiel, à l'adresse suivante : http://bergmanlab.ls.manchester.ac.uk/text2genome/about.cgi.

[47] Voir Maximilian Haeussler, Martin Gerner, et Casey M. Bergman, « Annotating genes and genomes with DNA sequences extracted from biomedical articles », *Bioinformatic*, 2011, p. 980-986. (non paginé dans sa version électronique) (disponible en ligne : <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3065681/>) : « The problem of finding relevant articles for a particular locus is becoming more acute as researchers increasingly adopt high-throughput genomic technologies (microarrays, genome-wide association studies, high-throughput sequencing, etc.). These genome-wide approaches often generate low-level data on thousands of genes or genomic regions, the interpretation of which becomes much more valuable when integrated with previously published studies on individual loci. »

[48] Voir Fidelia Ibekwe-SanJuan, *Fouille de textes : méthodes, outils et applications*, Paris, Lavoisier, 2007, p. 311.

formulation canonique de ce type de démarche, si dans un article un lien a été établi entre A et B, et dans un autre un lien entre B et C, il est probable que A influence C également[49].

À l'origine, le travail de Don R. Swanson était conduit manuellement : selon la méthode élaborée par ce chercheur[50], il s'agissait d'abord de choisir des sujets d'intérêt, de lire des publications pour sélectionner des concepts intéressants liés, de formuler et de vérifier des hypothèses par des lectures supplémentaires. Mais il semble évident que la prolifération de textes scientifiques publiés impose de suppléer à la lecture humaine par l'automatisation du processus de recherche de liens entre concepts. Depuis les premières tentatives d'automatisation, menées par Swanson lui-même au cours des années 90, ce domaine particulier de la recherche en text mining, qualifié dans la littérature anglo-saxonne d'*association extraction*, connaît toujours un développement actif : à partir du résultat de premières analyses qui extraient les entités nommées des articles, un second programme se charge de détecter des associations entre les entités, ce qui se fait en examinant la régularité des cooccurences entre différentes entités nommées[51].

Ainsi, recourir à la fouille de texte est chose rendue nécessaire par les tendances structurelles de la littérature scientifique, particulièrement dans le domaine biomédical. On sait que le développement de ces techniques dans leurs relations avec les politiques des éditeurs scientifiques est l'objet d'une actualité brûlante, sur laquelle nous reviendrons. Il faut toutefois déjà préciser de quoi le text mining est le nom, puisque l'expression n'est pas sans présenter plusieurs difficultés comparables à celles de la notion de mégadonnées, dont le succès est allé de pair avec une certaine dilution du sens.

LE TEXT ET DATA MINING : UNE PRATIQUE AU CARREFOUR DE PLUSIEURS DISCIPLINES

L'ambiguïté du terme de fouille de texte était déjà relevée par Fidelia Ibekwe-Sanjuan, qui écrivait en 2007 que sous cette étiquette passaient désormais « à peu près tous les travaux mettant en œuvre des processus automatiques pour traiter l'information textuelle »[52]. Le text mining tend donc à devenir appellation générique, qualifiant tout type de traitement informatique de données textuelles, quand bien même les traitements que l'on englobe ainsi peuvent également être revendiqués par des traditions différentes et plus anciennes, comme la lexicométrie (ou statistique textuelle) ou le traitement automatique du langage naturel (TALN).

[49]Dans Fidelia Ibekwe-SanJuan, *Ibidem*, p. 311, l'auteur donne l'exemple suivant concernant les travaux de Swanson, la mise en évidence d'un lien de causalité entre le syndrome de Raynaud et l'huile de poisson. « A partir des notices de Medline, il a trouvé que cette maladie (sujet A) pouvait être traitée avec l'huile de poisson (sujet C) par le cheminement suivant : la maladie de Raynaud est aggravée par une viscosité élevée du sang (*high blood viscosity*), l'agrégabilité des plaquettes sanguines (*platelet aggregability*) et la vasoconstriction. Or ces trois effets sont atténués par l'huile de poisson » En se fondant uniquement sur l'observation de la littérature scientifique et la mise en relation d'éléments qui y étaient présents mais disjoints, Swanson formula en 1986 l'hypothèse que l'huile de poisson pouvait avoir une influence bénéfique dans le traitement du syndrome de Raynaud. En 1989, une équipe de cliniciens valida empiriquement cette hypothèse (voir : http://isdm.univ-tln.fr/PDF/isdm12/isdm12a109_pierret.pdf).

[50]Voir Fidelia Ibekwe-SanJuan, *Ibidem*, p. 312

[51]Le processus est expliqué dans Graciela H. Gonzalez, Tasnia Tahsin, Britton C. Goodale *et alii*, « Recent advances and emerging applications in text and data mining for biomedical discovery », *Briefings in bioinformatics*, 29 septembre 2015 (disponible en ligne sur : <http://bib.oxfordjournals.org/content/early/2015/09/28/bib.bbv087.full#ref-14>) (consulté le 5 décembre 2015).

[52]Fidelia Ibekwe-SanJuan, *Fouille de textes : méthodes, outils et applications*, Paris, Lavoisier, 2007, p. 36.

Cette annexion par la fouille de texte des disciplines qui lui préexistaient[53] n'est pas le moindre des problèmes soulevés par sa définition. Avant de revenir sur l'historique de la fouille de texte et sur ses spécificités par rapport à d'autres méthodes en informatique des corpus, fournissons dès à présent une première définition de la fouille telle qu'elle se présente désormais : la fouille de texte est un ensemble de techniques informatisées qui visent, par l'emploi d'outils d'analyse statistique et linguistique, à extraire de l'information des textes ou à en révéler des structures ou motifs sous-jacents. Elle intègre des techniques permettant la recherche d'information, la reconnaissance de concepts ou d'entités au sein des textes, ainsi que le classement automatisé des textes. Pour cette activité, elle peut avoir une dimension descriptive, quand elle répartit en différentes catégories des textes réunis par leurs caractéristiques lexicales, ou une dimension prédictive ; dans ce dernier cas, après avoir déterminé quelles sont les caractéristiques sur le plan de la statistique textuelle d'un ensemble de textes, une application typique de text mining analysera un nouveau texte pour déterminer à laquelle des catégories documentaires précédemment fixées le rattacher. La fouille de texte se trouve héritière de plusieurs branches de la recherche en informatique : la recherche d'information, l'intelligence artificielle, le traitement automatique du langage naturel.

Il va sans dire que cette technique prend tout son sens appliquée à de grandes collections de textes, ou à des collections qui connaissent un accroissement permanent et rapide – comme des flux de courriels ou de messages sur un réseau social -, pour suppléer à une impossible exhaustivité de la lecture humaine.

Nous avons signalé ci-dessus que le text mining devenait appellation générique pour toute opération informatisée d'analyse textuelle. Et en effet, les différences se sont estompées entre le Traitement automatique du langage naturel et le text mining, à tel point que d'éminents spécialistes du TALN déclarent ne plus y voir de différences[54]. Cependant, cette indifférenciation progressive s'est développée en partie parce que les techniques d'analyse ont circulé entre les différentes disciplines : au niveau des buts poursuivis, la différence est plus nette. Pour distinguer le text mining d'autres types d'analyses informatisées de corpus, il est possible de pointer les différences dans les objectifs : alors que le traitement automatisé du langage naturel vise à appréhender la signification des contenus textuels[55], que la statistique textuelle (ou textométrie[56]) cherche à mettre en

[53]En effet, les spécialistes en statistique textuelle ou lexicométrie font remonter l'origine de la discipline à la décennie 1970 (voir Bénédicte Pincemin, Serge Heiden, « Qu'est-ce que la textométrie ? » (non paginé), disponible en ligne : <http://textometrie.ens-lyon.fr/spip.php?rubrique80>, consulté le 5 décembre 2015). Le TALN, pour sa part, remonte à des recherches sur la compréhension par les ordinateurs du langage humain initiées dès les années 1950 (Voir Article « Traitement automatique du langage », *Wikipedia*, disponible en ligne : <*https*://fr.wikipedia.org/wiki/Traitement_automatique_du_langage_naturel>, consulté le 5 décembre 2015).

[54]Lors d'un entretien téléphonique, le professeur Jean-Marie Pierrel, professeur d'informatique à l'Université de Lorraine et chargé de mission sur le projet ISTEX , nous a confirmé ne pas voir de différences entre les tâches du text mining et celles du TALN. C'est semble t-il l'opinion du professeur Isabelle Le Tellier, également spécialiste en ingénierie linguistique, qui voit dans dans la fouille de texte une nouvelle appellation du Traitement automatisée du langage naturel, après l'addition de nouvelles tâches informatiques venues des premières applications de text mining. Voir Jacqueline Léon, Isabelle Le Tellier, Le data turn. Des premiers traitements statistiques du langage (1950-60) à la fouille de texte

[55] Le traitement automatique du langage naturel (TALN) est un ensemble de techniques de traitement informatique des textes qui se situe dans le prolongement de la recherche en linguistique. Historiquement, la recherche en TALN s'est structurée autour de deux problèmes majeurs, qui sont la traduction automatique et la compréhension par les ordinateurs du langage humain. Voir l'article « Traitement automatique du langage naturel » de *Wikipedia* [En ligne]. Disponible sur https://fr.wikipedia.org/wiki/Traitement_automatique_du_langage_naturel. Consulté le 10 décembre 2015.

[56]La textométrie est le nom français de l'analyse statistique textuelle, qui s'est développée à partir des années 70. Voir PINCEMIN Bénédicte, HEIDEN Serge (2008) - "Qu'est-ce que la textométrie ? Présentation", *Site du projet Textométrie* , http://textometrie.ens-lyon.fr/spip.php?rubrique80

évidence des faits lexicaux ou syntaxiques en vue de mieux interpréter les textes d'entrée (c'est donc une aide à l'interprétation des corpus), la fouille de texte se propose d'extraire d'une collection de textes une information originale, et de préférence utile[57]. Cette définition ambitieuse, qui justifie la métaphore minière, a pu être contestée[58], dans la mesure où cette extraction semble un idéal rarement atteint. Elle met pourtant utilement l'accent sur le caractère pragmatique de cette activité, qui s'impose avec la force de l'évidence quand on considère l'étendue de ses domaines d'applications par rapport aux finalités, plus spécialisées et scientifiques, historiquement assignées au TALN ou à l'analyse textuelle : le text mining, outre ses apports déjà signalés pour la recherche scientifique, est présent par exemple dans le secteur commercial, où il sert dans le domaine de la relation avec le client, où il peut être utilisé pour la redirection automatique de courriels ou la catégorisation de courriels. Chacun a pu faire l'expérience d'au moins une application de text mining en consultant sa messagerie de courriels : les détecteurs de pourriels sont fondées sur des outils de classification automatique des courriels, qui analysent les courriers rejetées par l'utilisateur.

Tard-venue, l'expression de fouille de texte semble s'être imposée comme désignation générique pour l'analyse textuelle par l'étendue de ses applications ainsi que par sa diffusion dans de larges pans de la société, qui expliquent que d'autres écoles, avec qui elle partage certains outils statistiques et linguistiques, aient fini par se ranger sous sa dénomination. Ce qui ne signifie pas que les différences soient inexistantes : dans son *Introduction à la fouille de texte*[59] Isabelle Le Tellier propose une distinction entre les outils du TALN, coûteux en développement, utilisant des méthodes héritées de l'intelligence artificielle classique, dédiées à la compréhension des langues en analysant les structures syntaxiques des textes, et les outils de la fouille de texte, très bien adaptés à l'extraction d'informations de fortes quantités de textes mal structurés, tels que ceux que le développement d'Internet a permis de produire et de mettre en ligne. Nous rejoignons une distinction déjà évoquée précédemment : la fouille de texte cherche à extraire de l'information, à partir de quelques tâches simples, sans nécessairement comprendre le sens des textes, quand la mission historique du TALN était bien leur compréhension par des méthodes d'analyse linguistique.

Historiquement, on peut d'abord rappeler que la fouille de texte dérive d'abord de la fouille de données (data mining), discipline plus ancienne dont la fouille de texte emprunte le nom : celle-ci s'appliquait avant tout aux données structurées résidant dans des bases de données[60], et faisait partie d'une discipline, l'Extraction de connaissance dans des bases de données, qui avait pour but de

[57] Voir l'article suivant : http://bib.oxfordjournals.org/content/early/2015/09/28/bib.bbv087.full#ref-27 « Text mining extracts information from within those documents and aggregates the extracted pieces over the entire collection of source documents to uncover or derive new information. This is the preferred view of the field that allows one to distinguish *text mining* from *natural language processing* (NLP) » Cette définition rappelle une autre définition proposée par Marti Hearst, et citée dans Fidelia Ibekwe-SanJuan, *Ibidem*, p. 34 : la « fouille de texte » serait « la découverte par l'ordinateur de nouvelles informations extraites de plusieurs sources, qui seront reliées pour former de nouvelles réalités ou de nouvelles hypothèses à explorer par d'autres moyens. » Cette définition, exigeante, met l'accent sur le fait que la fouille de texte devrait permettre la découverte de motifs inconnus de l'utilisateur, de véritables « pépites de connaissances », ce en quoi le text mining se distinguerait clairement de la recherche d'information. Soulignons toutefois que cette définition est considérée comme trop restrictive par Fidelia Ibekwe-SanJuan, et qu'elle ne tient pas compte en effet de tous les cas de figures.

[58] Voir Fidelia Ibekwe-SanJuan, *Ibidem*, p. 37 qui souligne que l'extraction de nouvelles connaissances inconnues des textes est un objectif trop ambitieux, et préfère une définition de la fouille plus modeste, centrée sur l'extraction d'informations.

[59] Isabelle Le Tellier, *Introduction à la fouille de texte*, Paris, Université de Paris 3 – Sorbonne Nouvelle, (non paginé) (disponible en ligne sur : <http://www.lattice.cnrs.fr/sites/itellier/poly_fouille_textes/fouille-textes.pdf>) (consulté le 20 novembre 2015).

[60] Voir Fidelia Ibekwe-SanJuan, *Ibidem*, p.24.

« découvrir des informations implicites, inconnues jusqu'alors et potentiellement utiles et compréhensibles à partir des données »[61]. Par informations implicites, inconnues jusqu'alors, on entend essentiellement des motifs statistiques ou des structures dans les données. Les premières expériences de fouille de texte ont donc consisté en l'application des mêmes méthodes statistiques à ce matériau qu'était le texte, sans prendre en considération ses spécificités linguistiques ou sémantiques[62].

Ce bref retour sur les origines de la fouille de texte ne suffit pas à donner une vision minimale des tâches qu'englobe cette discipline en évolution rapide. Il faut donc reprendre en détail les tâches traitées par la fouille, et pour cela, on reprendra la répartition proposée par Isabelle Le Tellier, qui distingue quatre tâches majeures de text mining[63] : la recherche d'information, la classification, l'annotation, et l'extraction d'informations. Plus haut, un exemple typique de recherche d'information a été décrit avec le projet Text2Genom, mais on peut considérer que les grands moteurs de recherche sur Internet sont également des exemples évidents.

La classification est la classification automatique de documents : des textes partageant des caractéristiques communes sur le plan lexical, dévoilées par l'analyse statistique seront ainsi classés dans une même catégorie. L'application la plus répandue du procédé reste les programmes de classifications automatiques installées dans les logiciels de messagerie, qui détectent les courriers indésirables.

L'annotation est le procédé qui consiste à étiqueter chaque élément d'un texte, le plus souvent en spécifiant pour chaque mot sa catégorie morpho-syntaxique. Intégré dans un programme d'extraction d'information, l'étiquetage peut servir de phase préparatoire au repérage des informations à extraire.

L'extraction d'information consiste à extraire automatiquement des textes des informations où des éléments. Une application de cette voie qui connaît un développement rapide est la reconnaissance d'entités nommées : il s'agit de reconnaître des noms de lieux, de personnes, ou des dates dans un texte. Mais dans le cadre de la bibliométrie, c'est également la reconnaissance d'entités nommées qui permet à un service comme Google Scholar de repérer automatiquement les citations.

Ces quatre tâches sont distinguées à des fins heuristiques, mais elles ne représentent pas des éléments inconciliables dans les programmes d'informatique : il est fréquent que des programmes complexes et ambitieux les associent.

Bien que le text mining ait vu le jour avec des objectifs très pragmatiques, signalons enfin que l'usage a consacré un emploi de l'expression qui l'identifie purement et simplement à l'analyse statistique textuelle, telle qu'on la pratique à des fins de recherche en sciences humaines et sociales. Tel chercheur en histoire des discours politiques pourra ainsi caractériser de text mining sa pratique, fondée sur une analyse des occurrences et co-occurences de mots dans le discours des hommes politiques, tout en s'inscrivant dans la lignée des chercheurs en textométrie française, dont l'histoire débute plusieurs décennies avant qu'il soit question de fouille de texte, et avec des objectifs très différents[64]. Cette

[61]Voir Fidelia Ibekwe-SanJuan, *Ibidem*, p.24.

[62] Voir Fidelia Ibekwe-SanJuan, *Ibidem*, p.24. : « Compte tenu des difficultés inhérentes au traitement automatique des langues, à savoir les difficultés d'une analyse syntaxique et sémantique profonde des phrases, ces premiers travaux avaient opté pour des traitements de surface, assez robustes, qui permettaient une application directe des techniques de fouille de données aux textes. Ainsi les textes n'étaient pas vraiment considérés dans leur dimension linguistique ».

[63]Isabelle Le Tellier, *Introduction à la fouille de texte*, Paris, Université de Paris 3 – Sorbonne Nouvelle, (non paginé) (disponible en ligne sur : <http://www.lattice.cnrs.fr/sites/itellier/poly_fouille_textes/fouille-textes.pdf>) (consulté le 20 novembre 2015).

[64]Dans le cadre du séminaire « Introduction aux humanités numériques » organisé par l'ENS de Lyon, la séance du 4 décembre 2015 intitulée « Humanités numériques, text mining, big data : quelles pratiques quotidiennes derrière

indifférenciation progressive ne s'est pas faite sans raison, car il existe entre les deux pratiques d'évidentes parentés méthodologiques, fondées sur l'emploi des mêmes outils statistiques, ce que soulignent parfois les praticiens[65] ; ainsi la classification automatique des textes, une des tâches spécifiques du text mining, est également pratiquée par les chercheurs en textométrie, qui peuvent recourir à des algorithmes comme l'analyse arborée Luong pour identifier des regroupements de textes liées par des affinités linguistiques ou lexicales.

Cette identification du text mining à l'analyse statistique, bien qu'elle ne soit pas erronée, est cependant trompeuse, car elle néglige l'extension que les anglo-saxons accordent à cette discipline. La fouille de texte hérite en effet de la discipline de la recherche d'information (*Information retrieval*), qui est l'une des tâches phares actuellement de la fouille de texte. Le text mining inclut donc non seulement des tâches fondées sur la pure statistique textuelle, mais également des processus de recherche.

ces concepts à la mode » a permis à M. Damon Mayaffre de présenter ses travaux en cours. Dans sa présentation, il semblait clair que le text mining était considéré par lui d'abord comme une méthode d'analyse statistiques des textes, puisque les fonctionnalités présentées du logiciel Hyperbase concernaient surtout le comptage d'occurrence de termes, et autres démarches d'analyse statistique, plutôt que de recherche d'informations.

[65]On pourra consulter à ce sujet le site du projet textométrie, qui donne plus de détail sur les distinctions entre textométrie et « text mining » : Bénédicte Pincemin, Serge Heiden, « Qu'est-ce que la textométrie ? » (non paginé), disponible en ligne : <http://textometrie.ens-lyon.fr/spip.php?rubrique80>, consulté le 5 décembre 2015

BIG DATA ET SCIENCES HUMAINES : QUELS SERVICES A OFFRIR AUX CHERCHEURS ?

La recherche en sciences humaines et sociales a connu ces dernières années un développement marqué de l'intérêt pour les analyses menées à l'échelle des données massives. Dans le monde anglo-saxon, cette tendance s'exprime à travers le programme de financement « Digging into data », qui depuis 2009 soutient des équipes de recherche en sciences humaines engagées dans des projet de traitement et d'analyse de données massives, certaines en provenance de centres d'archives[66], d'autres issues de bibliothèques numériques[67]. Les collections numérisées peuvent donc faire l'objet d'un nouveau type d'appropriation de la part des chercheurs, auquel les établissements peuvent choisir de répondre en mettant en place des dispositifs dédiés. Ainsi, la bibliothèque numérique américaine Hathi Trust a t-elle choisi d'adosser à ses fonds numériques une plate-forme dédiée à la constitution de corpus et à l'élaboration d'algorithme de fouille de texte.

Cette partie s'intéresse donc à l'analyse de données massives comme à un nouvel enjeu pour les bibliothèques numériques, qui sont amenées à mettre en place différents dispositifs pour favoriser ce type de démarche. Pour donner à voir de plus près le mouvement et les enjeux de ce processus, il convient tout d'abord de brosser un rapide panorama de la recherche en sciences humaines quand elle se sert des données massives, mais non sans faire preuve de prudence : chaque projet ayant ses objectifs propres servis par des méthodes d'analyse informatiques parfois élaborées ad hoc, il n'est pas question d'opérer une synthèse artificielle, et d'évoquer un domaine de recherche unifié dont l'intitulé officiel serait « l'analyse de données massives au service des SHS », qui emploierait toujours les mêmes outils informatiques, et dont il serait possible de dresser la généalogie comme d'un objet permanent dans le temps. Si l'on évoque ce champ de la recherche par une formule synthétique, il s'agit d'abord d'une commodité de langage.

Cependant, malgré cette mise en garde, il existe bien des démarches de la recherche qui convergent autour du traitement des données massives, se rencontrent lors des mêmes événements institutionnels, et tiennent un même discours théorique. Même s'il ne saurait être question d'établir une généalogie rigoureuse de ces pratiques, elles ne sont pas sans rapport avec des techniques et des tendances de la recherche ayant existé précédemment. Et puisque les bibliothèques sont d'abord concernées par les techniques d'analyse textuelle, c'est d'abord de ces démarches qu'il sera question. Dans ce domaine, il semble clair que les méthodes d'analyse massive des données textuelles actuellement en usage à l'heure du big data (Traitement automatique du langage, text mining, etc...) dérivent de champs de recherches déjà anciens, dont on peut faire l'historique.

[66]C'est le cas par exemple, du projet mené par des participants du challenge autour des archives criminelles du Old Bailey, dont une présentation se trouve à l'adresse suivante : <http://www.oldbaileyonline.org/> (consulté le 6 décembre 2015).

[67]Comme le projet « Digging by debating », concernant les collections de la bibliothèque numérique HathiTrust, sur lequel on trouvera plus d'informations sur le site qui lui est consacré : <http://diggingbydebating.org/>.

LES SCIENCES HUMAINES ET LES APPROCHES DE TYPE BIG DATA

Les analyses textuelles assistées par ordinateur en sciences humaines : des outils forgés bien avant le big data

Au risque d'énoncer une évidence, commençons par dire que le traitement informatisé de corpus textuels a déjà une longue histoire derrière lui, et que toutes les approches élaborées au cours de cette histoire peuvent s'appliquer aux collections numérisées, à une échelle qui justifierait l'appellation d'approche « big data ». Ces méthodes « historiques » de l'analyse informatique de corpus peuvent relever de plusieurs approches ou traditions universitaire : nous avons déjà évoqué la statistique textuelle (ou encore appelée « textométrie » en France), qui applique des méthodes statistique aux corpus textuels pour en tirer des enseignements sur le plan stylistique, ou encore statuer sur l'attribution des textes à des auteurs sur la base de similitudes stylistiques quantifiées. On peut faire remonter en France l'émergence de ces méthodes aux années 1970, avec les travaux de Pierre Guiraud et de Charles Muller[68].

La linguistique informatique, dont on a vu que l'une des branches, le TALN, partage méthodes et outils avec la fouille de texte, débute au début des années 1950 avec des recherches sur la traduction des langues humaines [69]. L'utilisation de corpus en linguistique informatique commence, elle au cours des années 80, après la constitution des premiers grands corpus, tels que le corpus Brown aux États-Unis.

Les acquis de cette histoire longue de l'analyse textuelle sont aujourd'hui réinvestis dans des projets d'analyse qui tirent profit de l'abondance nouvelle de corpus textuels.

Les sciences humaines et le big data : de la réflexion théorique...

On voit que l'analyse quantitative assistée par ordinateur en sciences humaines a préexisté à l'invention même du concept de mégadonnées, et dans une situation de relative pénurie de corpus informatisés. Pour des raisons évidentes, l'avènement des grandes bibliothèques numériques, en mettant un terme à cette situation de rareté des textes informatisés, va impulser un renouveau de ces approches. Aux sources du « Digging into data challenge », le programme de subvention qui dans le monde anglo-saxon structure la recherche en sciences humaines sur les mégadonnées, on trouve par exemple un article intitulé « Que faire d'un million de livres », rédigé par le professeur Gregory Cane[70].

Mais ce nouveau souffle donné aux analyse de grands ensembles de données textuelles n'est pas la simple reconduction des projets scientifiques préexistants, désormais appliqués à des corpus plus étendus : les mégadonnées ont donné naissance à des projets scientifiques originaux, dont les ambitions dépassent les objectifs, plus circonscrits, des projets précédents en analyse textuelle. Pour les

[68] Bénédicte Pincemin, Serge Heiden, « Qu'est-ce que la textométrie ? » (non paginé) (disponible en ligne : <http://textometrie.ens-lyon.fr/spip.php?rubrique80>) (consulté le 5 décembre 2015).

[69] Article « Linguistique informatique », *Wikipedia* (disponible en ligne sur : <https://fr.wikipedia.org/wiki/Linguistique_informatique>) (consulté le 5 décembre 2015).

[70] Gregory Cane, « What do you do with a million books », *D-Lib Magazine*, Mars 2006 (Disponible en ligne : http://www.dlib.org/dlib/march06/crane/03crane.html) (consulté le 6 novembre 2015). Le rôle de déclencheur de cet article dans la mise en place du « Digging into data challenge » est mis en avant dans Christa Williford, Charles Henry, *One culture : computationnaly intensive research in the Humanities and Social sciences : a report*, p. 8

auteurs d'un article fameux[71] que nous devons citer, l'augmentation du nombre de textes numérisés disponibles permet désormais de formuler des hypothèses globales sur les courants de l'histoire culturelle:

> Lire des collections réduites d'œuvres soigneusement choisies permet au chercheur de construire des inférences puissantes au sujet des tendances de la pensée humaine. Cependant, cette approche ne permet que rarement une mesure précise des phénomènes sous-jacents. Les tentatives pour introduire les méthodes quantitatives dans l'étude de la culture ont été entravées par le manque de données convenables. Nous annonçons la création d'un corpus de 5 195 769 livres numérisés, corpus comprenant environ 4 % de tous les livres jamais publiés. L'analyse informatique de ce corpus nous permet d'observer des courants culturels et de les soumettre à des enquêtes quantitatives. Les études « culturomiques » élargissent les frontières de la recherche scientifique à un large spectre de nouveaux phénomènes[72].

La « culturomique » serait une nouvelle discipline vouée à l'étude de la culture humaine, à partir des grands réservoirs de données. Dans l'esprit des auteurs de l'article, ce nouveau domaine de la recherche devait trouver son outil d'élection dans l'application Google Ngram Viewer, dont ils sont également les concepteurs, et qui propose de visualiser par des graphiques les fréquences d'apparition des termes dans une partie du corpus des livres numérisés par Google books. L'application peut permettre ainsi d'effectuer des recherches sur la fréquence d'expressions comprenant de un à cinq mots, et de visualiser l'évolution dans le temps de ces fréquences d'apparition.

Ce programme de recherche a suscité au mieux le scepticisme, au pire les critiques acerbes, d'une partie de la communauté scientifique, qu'a laissée dubitative l'ambition d'analyser des mouvements culturels profonds par le simple comptage des occurrences de termes dans un corpus aussi hétérogène que Google Books[73] : on ne peut donc pas pour le moment parler d'un véritable engouement pour ce que Michel et Aiden ont appelé la culturomique. Cependant, ce texte nous intéresse car s'y lisent les nouvelles ambitions que la numérisation massives des collections suscite chez les praticiens de techniques d'analyse statistique, qui, en elles mêmes ne sont pas récentes. Auparavant, la statistique textuelle et le traitement automatisée du langage restreignaient leurs ambitions à la découverte des secrets que leurs corpus réduits pouvait livrer. Pour les personnes qui ont créé

[71] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden *et alii*, « Quantitative analysis of culture using millions of digitized books », *Sciencexpress*, 16 décembre 2010 (disponible en ligne sur : <http://www.librarian.net/wp-content/uploads/science-googlelabs.pdf>) (consulté le 16 décembre 2015).

[72] Voir Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser, *Ibidem*, p. 1 : « Reading small collections of carefully chosen works enables scholars to make powerful inferences about trends in human thought. However, this approach rarely enables precise measurement of the underlying phenomena. Attempts to introduce quantitative methods into the study of culture (1-6) have been hampered by the lack of suitable data. We report the creation of a corpus of 5,195,769 digitized books containing ~4% of all books ever published. Computational analysis of this corpus enables us to observe cultural trends and subject them to quantitative investigation. 'Culturomics' extends the boundaries of scientific inquiry to a wide array of new phenomena. »

[73] Certaines raisons de suspecter la pertinence des résultats obtenus par Ngram Viewers sont évidentes : puisque les graphiques sont élaborés à partir d'un dénombrement simple des occurrences de mots dans le corpus, sans que soit prise en compte la polysémie du vocabulaire, il n'est pas certain que tous les termes comptés fassent référence au même concept. Comme le disent F. Chateauraynaud et Josquin Debaz, cela « suppose de faire confiance aux mots, à la stabilité de leur sémantique, pour extraire des faits interprétables, sur une durée tellement longue qu'il ne soit pas garanti qu'un chat soit encore un chat. » Voir Francis Chateauraynaud, Josquin Debaz, « Prodiges et vertiges de la lexicométrie », *Socio-informatique et argumentation* (blog), 23 décembre 2010 (disponible en ligne sur : <http://socioargu.hypotheses.org/1963>) (consulté le 16 décembre 2015).

le néologisme de « culturomics », la réunion de la statistique textuelle et du développement d'immenses collections numérisées permet désormais de formuler des hypothèses sur l'évolution des cultures, et non seulement sur la linguistique ou la lexicologie.

Le recul manque encore pour apprécier les résultats de ce programme de recherche, aussi cet exemple est-il cité moins pour les résultats portés à son actif que comme le révélateur d'un changement d'ambitions des approches quantitatives à l'ère des données massives. On trouve des tendances comparables à l'œuvre également dans le champ plus restreint de la critique littéraire, où l'analyse informatique de très grands corpus est créditée par Matthew L. Jockers d'un grand mérite, celui de ne plus faire reposer les hypothèses en histoire littéraire sur des corpus trop restreints. Les analyses de mégadonnées permettraient donc aux études littéraires d'affermir leur assise scientifique[74], en dépassant l'analyse du petit nombre d'œuvres que le spécialiste pouvait lire in extenso. C'est donc l'espoir d'un véritable renouvellement épistémologiques qui se lit dans des lignes comme les suivantes :

> Aujourd'hui […] l'ubiquité des données, ce que l'on appelle les mégadonnées, change la pratique de l'échantillonnage. De fait, les mégadonnées sont en train de modifier les fondements de la pratique des sciences humaines. L'existence d'énormes entrepôts de données signifie que de nombreux domaines de recherche ne dépendent plus d'expériences contrôlées et artificielles, ou d'observations dérivées. Plutôt que de conduire des expériences contrôlées sur des échantillons et d'en extrapoler du spécifique au général ou du proche au lointain, ces immenses jeux de données permettent des recherches à une échelle qui atteint ou approche l'exhaustivité. La « population », autrefois inaccessible, est devenue à portée et remplace rapidement les échantillons aléatoires et représentatifs.[...] Pour les études littéraires, nous avons l'équivalent de ces grands ensembles de données sous la forme de grandes bibliothèques[75].

Qu'il s'agisse d'analyser des phénomènes culturels, ou encore de formuler des hypothèses d'histoire littéraire, ces deux exemples montrent que ce qui est à l'œuvre dans les approches big data en sciences humaines, c'est souvent une tentative de dépassement des biais imposés par l'analyse attentive d'échantillons trop restreints. La dichotomie popularisée par le critique littéraire Franco Moretti entre la lecture rapprochée (*close reading*) et la lecture distante (*distant reading*) innerve ces débats, où l'analyse informatique de mégadonnées est souvent présentée comme le prolongement technique d'une méthodologie qui souhaite mieux rendre compte des mouvements de fond de l'histoire littéraire, en substituant

[74]Matthews L. Jockers ne peut cependant pas être accusé de scientisme : il dit par ailleurs que les résultats des analyses informatiques ne sont rien sans interprétation.

[75]Matthews L. Jocker, *Macroanalysis*, Springfield, University of Illinois Presse, 2013 p. 7 : « Today, however, the ubiquity of data, so-called big data, is changing the sampling game. Indeed, big data are fundamentally altering the way that much science and social science get done. The existence of huge data sets means that many areas of research are no longer dependant upon controlled, artificial experiments or upon observations derived from data sampling. Instead of conducting controlled experiments of samples and then extrapolating from the specific to the general or from the close to the distant, these massive data sets are allowing for investigations at a scale that reaches or approaches a point of being comprehensive. The once inaccessible "population" has become accessible and is fast replacing the random and representative sample. In literary studies, we have the equivalent of this big data in the form of big libraries. »

à la lecture minutieuse de quelques textes triés sur le volet l'analyse de corpus plus vastes, faisant l'objet d'une lecture non exhaustive et partielle. L'idée de lecture distante avait d'abord été formulée par Moretti dans le contexte spécifique des études littéraires, sans référence à des analyses informatiques, comme une méthode pour appréhender la « masse sans lecteurs » (*the great unread*) des textes ignorés par l'histoire littéraire classique [76] : il s'agissait alors, tout simplement, de promouvoir des approches alternatives à la lecture intégrale des œuvres, comme l'analyse des titres, pour prendre compte de plus d'œuvres que n'en contient le canon des textes reconnus comme classique. Cependant, cette formule heureuse de « lecture à distance » s'est trouvée, bien des années après, convenir parfaitement pour la démarche adoptée par les analystes en données massives ; aussi les praticiens des humanités numériques se la sont-elles appropriées, et il est tout à fait courant de voir la « lecture à distance » citée dans des articles sur le big data, qui tendent parfois à l'employer comme un synonyme des approches à l'échelle des mégadonnées. Franco Moretti lui-même s'est engagé dans l'analyse des données massives, puisqu'il dirige désormais le Litterary Lab de Stanford, département de recherche en histoire littéraire spécifiquement voué aux approches quantitatives assistées par ordinateur.

Analyser des phénomènes culturels auparavant inaperçus en raison de l'absence de données, ou affermir des hypothèses en sciences humaines par l'analyse informatisée de corpus immenses que la lecture humaine ne saurait appréhender, telles sont deux des promesses les plus intéressantes de l'avènement des données massives en sciences humaines.

… à la pratique : l'atelier des humanités numériques à l'âge des mégadonnées

Quelque enthousiasmant que soit le renouvellement épistémologique que les mégadonnées laissent entrevoir, on ne prendra vraiment la mesure du mouvement pour les sciences humaines qu'après avoir examiné par le menu quelques exemples de projet de recherche. Pour ce faire, on ne peut faire l'économie d'évoquer le principal relais institutionnel qui structure ce champ de la recherche dans le monde anglo-saxon, le challenge « digging into data »[77].

Organisé depuis 2009 sous la forme d'un concours, il permet à chacune de ses sessions de subventionner plusieurs projets de recherche, dont tous ont en commun l'analyse informatique de données diverses, caractérisées par leur volume et leur hétérogénéité, ainsi que de concerner la recherche en sciences humaines et sociales. Il s'agit, à l'origine, d'une initiative portée par l' « Office of digital humanities » du « National endowment for the humanities », l'agence fédérale américaine chargée de la promotion des projets d'humanités numériques. Le projet a d'abord été porté avec l'espoir de faire émerger une culture commune à l'informatique et aux sciences humaines. Rapidement, l'initiative américaine devient internationale, puisque le Joint Information Systems Committee de Grande-Bretagne et le Canadian Social Sciences and Humanities Research Council rejoignent l'initiative.

[76]Pour les détails du cheminement intellectuel de F. Moretti, on se reportera à Franco Moretti, *Distant reading*, New-York, Verso, 2013, et pour la genèse de la « lecture distante », plus particulièrement aux articles « *Conjectures on World Literature* » et « *The Slaughterhouse of Literature* » de ce recueil.

[77] Pour plus d'informations, on pourra se reporter au site officiel de cette manifestation : http://diggingintodata.org/.

Les projets soutenus par ce système de subvention permettent un premier aperçu sur les méthodes et les pratiques des chercheurs en sciences humaines, quand ils travaillent sur les mégadonnées. On s'en tiendra ici à des exemples de recherches concernant des ensembles de données textuelles, comme les plus proches des fonds de bibliothèques, dont certaines ont fourni les données numériques sur lesquelles travaillaient les chercheurs.

Ainsi, l'une des équipes de recherche lauréates a pris pour base de son travail deux millions six cent mille livres des fonds numérisés de la bibliothèque numérique Hathi Trust, dans un projet visant à analyser les points de contacts entre la philosophie et les sciences, par identification des arguments de nature philosophique dans des textes scientifiques[78] : l'une des premières étapes du projet consistait à utiliser un algorithme de fouille de texte, pour modéliser les thèmes abordés dans les textes, dans un travail dont le but était de cartographier les arguments utilisés au sein des textes.

Les travaux d'un autre groupe de recherche attirent l'attention, car concernant un type de documents massivement présents dans les fonds des bibliothèques, à savoir les collections de périodiques : sur un corpus de plus de cent mille articles de journaux couvrant les années 1917 et 1918, l'application par une équipe de chercheurs américains d'un programme de fouille de texte a accompagné l'analyse de la réponse sanitaire face à la propagation aux Etats-Unis de la grippe espagnole, dans un projet dont l'objectif était d'étudier la façon dont les journaux influençaient l'opinion publique et diffusaient les discours autorisés [79]. Pour étudier la propagation des nouvelles épidémiologique, les chercheurs ont utilisé une méthode particulière de fouille de texte, la modélisation de sujet (*topic modeling)*, qui leur a permis d'examiner quels termes étaient fréquemment employées ensemble dans les premières nouvelles concernant la grippe.

Les sciences humaines à l'échelle des données massives ne sont bien entendues pas l'apanage du seul monde anglo-saxon, et des projets similaires se trouvent également en France, quoique dans un contexte juridique plus incertain pour la fouille de texte. Le laboratoire OBVIL (Observatoire de la vie littéraire) a ainsi entrepris un projet de recherche visant à repérer dans un corpus de textes littéraires les réseaux de citations entre textes : ce qui passe par l'emploi de procédures de fouilles de textes pour repérer automatiquement, au sein du corpus, les textes qui sont cités par d'autres[80].

LES DISPOSITIFS MIS EN PLACE EN BIBLIOTHEQUES NUMERIQUES

Pour répondre au développement de ces usages originaux, les bibliothèques peuvent mettre en place plusieurs actions : si certaines s'inscrivent dans la continuité de leurs domaines d'expertise traditionnels, en se centrant sur l'administration des documents numériques, dont les caractéristiques doivent

[78]Voir Simon McAlister, Colin Allen, Andrew Ravenscroft, Chris Reed, David Bourget, John Lawrence, Katy Börner et Robert Light, *From big data to argument analysis and automated extraction : a selective study of argument in the philosophy of animal psychology from the volume of the Hathi Trust collection*, 2014, p. 1 (disponible sur le site : http://diggingbydebating.org/wp-content/uploads/2014/04/DiggingbyDebating-FinalReport2.pdf) (consulté le 5 décembre 2015).

[79]Pour une présentation de ce projet, voir : Tom Ewing, *Data mining the 1918 Influenza Pandemic : an introduction to the epidemiology of information*, 2012 (disponible en ligne : <http://www.flu1918.lib.vt.edu/wp-content/uploads/2012/11/VT-DiD-Presentation-JCDL-12June2012.pdf>) (consulté le 20 décembre 2015). Le projet a également mis en ligne un site Internet : <http://www.flu1918.lib.vt.edu/>.

[80]Information recueillie au cours d'un entretien avec M. Jean-Gabriel Ganascia, directeur adjoint du LABEX OBVIL.

permettre les exploitations informatiques, d'autres paraissent plus inédites, comme la mise à disposition par les bibliothèques des moyens informatiques (ressources en calcul) pour effectuer des algorithmes de fouille de texte sur les corpus hébergés. La bibliothèque, dès lors, n'est plus seulement entrepôt documentaire, mais contribue à l'analyse des collections par les instruments qu'elle fournit, ou par l'assistance qu'elle peut apporter à l'analyse des collections.

Empressons nous d'ailleurs de préciser que les deux approches sont loin d'être exclusives l'une de l'autre, puisqu'une bibliothèque numérique proposant une plate-forme d'analyse informatique devra, cependant, s'assurer que les textes mis à disposition sont bien correctement formatés pour permettre le TDM.

Un préalable indispensable : l'administration de corpus documentaires d'une taille critique, et adéquatement formatés pour la fouille de texte

La mise à disposition de corpus importants et convenablement formatés pour des opérations de fouille semble bien le premier travail à effectuer par les bibliothèques pour permettre des analyses de type big data. Revenons d'abord sur la question du volume : ici, elle s'entend à la fois sous l'angle restreint de la taille documentaire, mais également comme un impératif de dépassement des divisions introduites par les éditeurs dans les résultats de la recherche scientifique. Plus haut, quand la question de la définition des mégadonnées a été abordée, la « recherche de signaux faibles » a été présentée comme le trait marquant des approches qui caractérisent le big data comme une méthodologie de l'analyse informatique. Dans cette approche, c'est bien la taille considérable des corpus de données qui permet de percevoir à l'intérieur de ces ensembles des corrélations faibles, donc de faible ampleur, qui ne seraient pas perceptibles sur des volumes plus réduits de données. Ainsi, en ce qui concerne le TDM, la quantité de données analysées influe la qualité de la recherche, en rendant perceptibles des phénomènes différents.

En ce qui concerne cet intérêt du volume pour le TDM sur les articles scientifiques, le morcellement des publications à travers différents entrepôts de données, correspondant aux collections d'éditeurs différents, pose de façon évidente un problème. De cette façon, le paysage éditorial risque d'influer sur la recherche, en rendant difficile l'extraction automatisée d'informations sur l'ensemble des publications concernant un domaine de la recherche, puisque les articles sont répartis arbitrairement à travers les collections des éditeurs, dans des silos de données séparés.

La première réponse proposée pour pallier cet état de fait est la constitution d'entrepôts rassemblant cette production éparse, et ce afin de permettre des opérations d'extraction automatisée de connaissances sur des corpus qui transcendent les divisions du paysage éditorial. Il s'agit de l'un des enjeux, en France, du projet ISTEX (Initiative d'excellence en information scientifique et technique), qui rassemble un programme d'acquisitions de corpus documentaires par licences nationales passées avec les éditeurs, et la constitution d'une plate-forme qui rassemblera les corpus acquis, et proposera à ses usagers des services de base ainsi que des services à valeur ajoutée. L'ambition est donc de rassembler sur une même plate-forme des corpus pleins textes issus d'ensembles éditoriaux disjoints, et de permettre, sur l'ensemble documentaire ainsi constitué, le développement de nouveaux usages de recherche que permettra la taille du volume

constitué. La fouille de texte a été évoquée dès l'origine du projet comme le principal de ses usages nouveaux que favorisera la taille exceptionnelle du corpus[81].

Cependant, le projet ISTEX présente une autre caractéristique, celle d'inclure dans son action une importante phase de retraitement, de normalisation et d'enrichissement des textes et de leurs métadonnées. Ce second volet de l'activité du projet permet d'évoquer ce qui sera, sans doute, le rôle majeur des gestionnaires de grands corpus numériques pour le développement du TDM : veiller à la qualité les données textuelles, dans la perspective de permettre leur exploitation automatisée. En mars 2015, lors d'un séminaire organisé sur les questions techniques dans ISTEX, la faible qualité des textes fournis par les éditeurs et de leurs métadonnées était pointée comme l'une des principales difficultés du projet[82], constituant un obstacle majeur pour le développement des services innovants projetés pour la fouille de texte. . C'est un autre aspect du rôle des bibliothécaires dans le développement des usages de fouille de texte : à la constitution de vastes corpus de textes s'ajoute la curation des données pour permettre leur exploitation de façon automatisée.

Pour triviale que puisse paraître la remarque, le premier service que les bibliothèques numériques peuvent rendre aux chercheurs dans ce domaine demeure la mise à disposition de corpus de textes formatés de façon adéquate pour rendre l'analyse automatisée possible. Il est probable que ce problème restera, longtemps, une priorité, avec la clarification du statut juridique du text mining. Car si l'essor des grandes entreprises de numérisation est bien en partie à l'origine des projets de recherche sur les mégadonnées, il s'en faut de beaucoup pour que la seule diffusions des textes numérisés sur Internet suffise à permettre leur exploitation à grande échelle. Par exemple, les pratiques des chercheurs peuvent nécessiter que les textes se présentent autrement que sous la forme d'un simple fichier texte, en faisant également l'objet d'une structuration sémantique. Dans un article incisif sur les possibilités de fouiller la bibliothèque numérique Gallica [83], Pierre-Carl Langlais a ainsi listé les paramètres qui entrent en ligne de compte pour évaluer la compatibilité d'une bibliothèque numérique avec le TDM : l'accessibilité des fichiers, la lisibilité des url, la sémantisation des documents, la qualité des métadonnées et des données associées, le statut légal. À cette liste, nous pouvons, sans risque, ajouter la qualité de l'océrisation des fichiers, qui conditionne de façon évidente l'exploitation des textes.

Ces différents critères pour apprécier la capacité d'automatiser l'analyse de corpus, nous pourrions les répartir entre ceux qui représentent une nécessité, d'ordre à la fois technique et scientifique, pour permettre le TDM, et ceux demandés par les chercheurs pour faciliter cette opération. Dans la première catégorie, disposer de textes non erronés, sans lacunes, est la première condition – évidente - pour que des analyses de type big data soient menées. Mais cette

[81]Pour s'en convaincre, on peut consulter la synthèse de Raymond Berard, *ISTEX, vers des services innovants d'accès à la connaissance*, ABES, Montpellier, 2012, (Disponible en ligne : http://www.abes.fr/Ressources-electroniques2/Acquisitions/Licences-nationales-ISTEX) où on lit page 2 : « Le corpus résultant permettra par son importance la fouille de texte et pourra faire surgir de nouvelles pistes pour la recherche. » On lit également, page 1, « L'ensemble de données ainsi constitué sera disponible en permanence pour une ingénierie scientifique d'un niveau sans commune mesure avec celle des communautés qui travaillent aujourd'hui sur des plateformes juxtaposées »

[82]Voir le compte-rendu *Séminaire techniques services ISTEX*, mars 2015, p. 60 : « Le besoin d'améliorer les données et métadonnées d'ISTEX est prédominent dans les travaux menés par le développement des services de la plateforme ». Les membres du séminaire pointaient notamment la faible qualité du plein texte, obtenu par une océrisation de mauvaise qualité.

[83]Pierre-Carl Langlais, « Peut-on faire du data-mining sur Gallica ? », *Sciences communes*, 2014 (disponible en ligne : http://scoms.hypotheses.org/186) (Consulté le 6 décembre 2015)

condition suppose déjà un travail important de la part des gestionnaires de bibliothèques numériques, travail qui, dans bien des cas, dépend étroitement des progrès de la reconnaissance optique de caractères. Actuellement, on estime que les fichiers mal interprétés peuvent contenir jusqu'à 20 % d'erreurs[84], et, selon les valeurs du référentiel de la BNF, une conversion de très bonne qualité en mode texte d'un document numérisé peut avoir un taux de reconnaissance des caractères allant jusqu'à 99,9 %[85]. La vérification humaine est donc toujours nécessaire, et elle le sera toujours d'autant plus que, selon la formule d'Edwin Klijn, les erreurs d'une reconnaissance optique de caractères en disent parfois plus sur la qualité du matériau d'origine que sur la performance réelle du logiciel[86]. Le vieillissement du papier, les défauts d'impression, sont autant de qualité du matériau d'origine qui influent sur la conversion en mode plein texte : les taux de reconnaissance les plus bas sont attendus sur les collections de documents antérieurs à 1750. L'écriture humaine n'est pas actuellement reconnue par les logiciels disponibles, et ne le sera probablement pas avant longtemps, si elle finit par l'être. Cet ensemble de facteurs explique que, pour permettre la fouille de texte sur l'ensemble de leurs collections numériques, le premier soin des bibliothécaires devrait encore longtemps rester l'amélioration de la qualité des textes fournis, et donc la correction des fichiers obtenus par la conversion en mode texte. Des alternatives aux marchés de correction des OCR existent, à l'exemple de la correction par crowdsourcing, étudiée par M. Andro[87].

Toujours du côté des préalables indispensables, on citera l'administration de métadonnées de qualité. Comme la génération de métadonnées est souvent un processus automatisé, effectué en même temps que la reconnaissance optique de caractères des documents numérisés, les métadonnées peuvent souffrir des mêmes défauts, avec des conséquences importantes dans le cas d'analyse informatique massive de corpus, où un document mal daté pourra intégrer le champ d'une analyse qui ne le concerne pas. L'exemple de Ngram Viewer, de ce point de vue, est édifiant, puisque cette application fonctionne à partir des métadonnées de Google Books, de qualité aléatoire, notamment au niveau des dates de document : c'est pourquoi le résultat d'une recherche sur le bigramme « Barack Obama » semblera indiquer que des publications ont commencé à concerner le président américain l'année même de sa naissance[88]. Si cet exemple est trivial, on voit combien des métadonnées de faible qualité peuvent affecter des recherches plus sérieuses sur les mégadonnées, notamment pour les études linguistiques[89].

Bénéficier de textes de bonnes qualités, accompagnés de métadonnées du même niveau, telles seraient les bases pour permettre une exploitation automatisée des corpus par les chercheurs. Si banal que paraisse la constatation, l'exemple des

[84] Mathieu Andro, Imad Saleh, *La correction participative de l'OCR : le crowdsourcing au profit des bibliothèques numériques* (Disponible en ligne : http://bbf.enssib.fr/contributions/la-correction-participative-de-l-ocr#notes) (consulté le 7 décembre 2015)

[85] Bibliothèque national de France, *Référentiel OCR*, 2013 (disponible en ligne : <http://www.bnf.fr/documents/ref_num_ocr.pdf>) (consulté le : 16 décembre 2015).

[86] Voir Edwin Klijn, « The current State-in-art in Newspaper Digitization : a market perspective », *D-Lib Magazine*, Vol. 14 Janvier 2008 : « Accuracy rates, on either word or character level, should not be considered as watertight performance indicators for OCR software. Usually the quality of the OCR texts says more about the condition of the original materials than it does about the performance of the OCR software. » (disponible en ligne sur : <http://www.dlib.org/dlib/january08/klijn/01klijn.html>) (consulté le 23 décembre 2015).

[87] Mathieu Andro, Imad Saleh, *La correction participative de l'OCR : le crowdsourcing au profit des bibliothèques numériques* (Disponible en ligne : http://bbf.enssib.fr/contributions/la-correction-participative-de-l-ocr#notes) (consulté le 7 décembre 2015)

[88] L'application renvoyait ce résultat encore le 20 décembre 2015.

[89] Voir Sarah Shang, « The Pitfalls of Using Google Ngram to Study Language », *Wired*, 12 novembre 2015 (disponible en ligne sur : <http://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/>) (consulté le 5 décembre 2015).

collections rassemblées pour la plate-forme ISTEX montre qu'à l'heure actuelle, obtenir ce niveau de qualité pour l'ensemble des collections numériques reste un chantier important pour les établissements. D'autres caractéristiques semblent demandées par les chercheurs que nous avons interrogés, ou dont nous avons lu les avis, qui ne relèvent pas du même degré de priorité, mais sont cependant importantes. D'une part, les chercheurs relèvent l'intérêt de bénéficier d'un format qui soit structuré sémantiquement. Par sémantisation, on entend ici l'encodage des caractéristiques structurelles et sémantiques des textes : il est, par exemple, particulièrement intéressant de connaître si une chaîne de caractères appartient à un titre intermédiaire ou au corps du texte. Dans le cadre d'une recherche portant sur la fréquence d'utilisation de certaines catégories lexicales ou grammaticales dans les articles scientifiques, il est intéressant de savoir si l'endroit où la chaîne de caractère concerné apparaît est un titre ou dans le corps d'un paragraphe[90].

Faute de permettre un service permettant de lancer, directement sur les serveurs de la bibliothèque numériques, des algorithmes de fouille – comme nous allons voir qu'il en existe -, permettre une extraction aisée des corpus libres de droit et de leurs métadonnées semble une demande également importante de la part des chercheurs. L'un d'entre eux [91] souligne la difficulté d'opérer de telles extractions pour construire des corpus d'une taille suffisante pour étudier l'évolution de thématiques sur le long terme. Le besoin d'une API, c'est à dire d'une interface permettant l'extraction automatisée, est également soulignée[92] comme une nécessité pour faciliter la récupération des textes qui constituent la matière première des recherches.

Fournir aux chercheurs une plate-forme d'expérimentation

Cette idée de fournir aux chercheurs, en sus des indispensables corpus de qualité, des moyens techniques pour les analyser, a connu des réalisations rares, mais ambitieuses, aux États-Unis, où l'on peut citer quelques exemples de plate-formes (Hathi Trust Research Center et Jstor Data for research program) permettant aux chercheurs d'effectuer directement des opérations algorithmiques sur des corpus qu'ils constituent eux-mêmes à partir des collections. Ces réalisations semblent faire écho aux réflexions d'Eric Lease Morgan, bibliothécaire à l'université Notre-Dame, sur les services qu'un usager peut attendre d'une bibliothèque concernant les collections numériques : après avoir fourni l'accès au document, et les outils qui en facilitent la recherche, la prochaine étape dans l'évolution des bibliothèques numériques serait la fourniture aux usagers de services facilitant la compréhension et l'analyse des documents, par l'entremise de la fouille de texte[93].

Si les projets de recherche les plus ambitieux pour le traitement des mégadonnées nécessitent l'écriture d'algorithme de fouille de texte *ad hoc*, produit de développements sophistiqués, il existe également ce qu'on pourrait appeler un

[90]Exemple suggéré par M. Jean-Marie Pierrel, lors de l'entretien qu'il nous a aimablement accordé.

[91]M. Pierre Ratinaud, que nous remercions d'avoir bien voulu répondre à nos questions par courriel.

[92]Surtout par Pierre-Carl Langlais, « Peut-on faire du data-mining sur Gallica ? », *Sciences communes*, 2014 (disponible en ligne : http://scoms.hypotheses.org/186) (Consulté le 6 décembre 2015)

[93]Voir Eric Lease Morgan, *Use and understand : the inclusion of services against texts in library catalogs and « discovery systems »*, Library Hi-Tech, Vol. 30, 2012 p. 35-59 "Use and understand is a next step in the information flow. It comes after find and get, and it is a process enabling the reader to better ask and answer questions of an entire collection, sub collection, or individual work. By applying digital humanities computing processes, specifically text mining and natural langage processing, the process of use and understand can be supported a library catalog or "discovery system".

fond commun de l'analyse informatique de corpus, constitué d'algorithmes ayant fait leur preuve, et prêts à être appliqués sur tout corpus qu'on leur présentera. C'est pourquoi une des réponses qu'il est possible d'apporter à ces nouveaux usages des chercheurs est de leur proposer une plate-forme informatique, qui leur permette de sélectionner des livres pour constituer un corpus, et ensuite permettre de lancer sur ce corpus prédéfini le traitement de leur choix. Des algorithmes pré-écrits sont proposés aux chercheurs, mais ils sont libres également d'écrire les leurs, et de les partager avec d'autres chercheurs. Cette option est celle qu'a retenu la plus grande des bibliothèques numériques américaines, HathiTrust, qui a joint à son dispositif d'origine une infrastructure spécifiquement dédiée à l'analyse informatique de corpus, le HathiTrust Research Center.

Ce dispositif se présente sous la forme d'un site rattaché à celui de la bibliothèque, où l'usager peut s'inscrire pour avoir accès aux services dédiés. À son niveau le plus simple d'utilisation, l'usager peut constituer son propre corpus de recherche en sélectionnant des textes de la bibliothèque, puis appliquer sur ce corpus personnalisé certains des algorithmes prédéterminés sur le site : extractions d'entités nommées, compteur de la fréquence de mots... Le processus de fouille sera effectué à distance par les ressources informatiques du HTRC, et l'usager aura accès au résultat de l'opération. Pour les plus aguerris d'entre les praticiens de l'analyse textuelle, le site fournit également un « bac à sable », où écrire son propre algorithme, avant de le lancer sur le corpus préselectionné.

Cependant, les services que propose le HTRC ne se limitent pas à cette offre de base : pour des projets nécessitant des moyens supérieurs en puissance de calcul, le HTRC peut également apporter l'aide d'un super-ordinateur[94], ainsi que son expertise pour le développement d'algorithme. Un travail de recherche portant sur plus d'un million de textes, mené par le professeur Vernon Burton sur les stéréotypes concernant les habitants du Sud des États-Unis a ainsi été rendu possible grâce à l'aide apportée par le HTRC[95].

Il faut signaler une caractéristique du fonctionnement du HTRC, qui intéresse les discussions sur le statut juridique du TDM : le dispositif a été pensé pour permettre d'effectuer des opérations de fouille de texte sur des contenus libres de droits, mais également, selon des modalités spécifiques, sur des textes encore protégés par le copyright. En effet, selon la jurisprudence américaine, que nous aborderons plus en détail ultérieurement, la fouille de texte représente un « usage raisonnable » (*fair use*) des contenus sous droit, c'est à dire un usage qui ne porte pas préjudice aux ayant-droits, pour autant que le chercheur puisse appliquer son algorithme de fouille de texte sur les œuvres protégées par le droit d'auteur sans pouvoir en effectuer une copie. Le HTRC a donc élaboré un dispositif qui puisse permettre un usage des contenus sous droit par la fouille de texte, c'est-à-dire un système informatique qui permette à un chercheur de lancer une opération de fouille de texte sur des corpus, tout en sécurisant l'accès aux données sous droit. Il s'agit d'une machine virtuelle, à laquelle l'usager se connecte : il ne peut ensuite avoir accès aux contenus sous droits qu'en activant le mode protégé (*secure mode*) de la machine virtuelle, qui lui permet d'avoir accès aux collections protégées sur la machine, mais l'empêche dès lors d'accéder au reste du réseau, et donc d'extraire de l'environnement de travail des copies de contenus protégés. Le chercheur peut

[94]Le HathiTrust Researc Center a été développé conjointement par l'Université de l'Indiana et l'Université de l'Illinois, qui disposent de plusieurs super-ordinateurs. Pour plus d'informations sur l'organisation du HTRC, voir la diapositive suivante disponible en ligne : <https://www.hathitrust.org/documents/HathiTrust-TRLN-20140723.pdf>

[95]Lance Farrel, *HathiTrust Research Center adds 5 billion pages to help scholars see farther* (disponible en ligne https://sciencenode.org/feature/hathitrust-research-center-adds-5-billion-pages-help-scholars-see-farther.php) (consulté le 5 décembre 2015).

alors effectuer ses opérations de TDM dans cet environnement clos, d'où il pourra seulement extraire le résultat de ses analyses par un canal sécurisé[96].

[96]Voir Beth Plale, Atul Prakash, et Robert McDonald, *The data capsule for non-consumptive research : final report*, p. 2 (disponible en ligne sur : <https://scholarworks.iu.edu/dspace/bitstream/handle/2022/19277/HTRCSloanReport_ScholarWorks.pdf?sequence=1&isAllowed=y>) (consulté le 25 décembre 2015).

LA PERSPECTIVE DE NOUVEAUX INSTRUMENTS DE RECHERCHE POUR LES COLLECTIONS ?

Ces méthodes que nous avons vues utilisées avec intérêt par des chercheurs en sciences humaines, il peut sembler séduisant de les appliquer aux collections numériques, non à des fins d'analyse, mais d'orientation des lecteurs dans les collections ; si la recherche « plein texte » - une des applications de la fouille de texte - est déjà utilisée dans de nombreuses bibliothèques numériques, on a vu que la fouille de texte comprenait nombre d'autres méthodes, dont certaines peuvent être utilisées à bon escient dans les fonds. Ce qui signifie les adapter pour élaborer de nouveaux instruments de recherche, qui superposeraient une nouvelle méthode d'accès aux documents aux procédures classiques de recherche, telles que la recherche par mots d'autorités ou l'utilisation d'index.

Eus égards à la richesse et à la vitalité de la fouille de texte comme champ de recherche, il ne peut s'agir ici que de présenter quelques exemples particulièrement frappants des possibilités issues de cette discipline pour élaborer des instruments de recherche innovants. Quoique modeste, cette investigation nous permettra cependant d'approcher d'un peu plus près le fonctionnement des techniques de TDM, et de lever un peu plus le voile sur ces pratiques des chercheurs en sciences humaines dont on cherche à estimer le potentiel pour la recherche dans les fonds numérisés ou nativement numériques. Enfin, ce chapitre n'a d'autre ambition que de proposer un aperçu des voies que pourrait ouvrir la fouille de texte pour le développement d'outils de recherche innovants, sans prétendre sur ce sujet émettre des avis catégoriques. Les exemples donnés, notamment concernant l'application de la « modélisation de sujet » (ou *topic modeling*), ne sont pas présentés comme des modèles de réalisation achevée – et l'on pointera au passage leurs limites et imperfections -, mais comme des projets intéressants pour les perspectives qu'ils ouvrent sur ce que pourrait proposer à l'avenir le TDM pour les bibliothèques numériques, notamment sur l'analyse sémantique des collections.

CARTOGRAPHIE DE LA CONNAISSANCE

Par instrument de recherche en bibliothèques numériques, on entend surtout d'abord un moteur de recherche. Cependant, bien que la recherche d'information soit l'une des applications de la fouille de texte, nous avons vu qu'elle n'était pas la seule, et que le TDM, dans ses définitions les plus ambitieuses, se proposait de permettre la découverte de nouvelles connaissances, quand des applications de recherche de base, telles qu'implémentées dans les catalogues, proposent d'abord de trouver les documents pertinents pour une recherche, sans proposer les moyens d'analyse permettant de parvenir à une connaissance originale. L'application d'instruments de fouille de texte aux collections semble donc promettre des outils qui permettent à la fois de rechercher des documents, mais également d'entamer une analyse des collections selon différentes perspectives, dans un contexte où les bibliothèques numériques doivent non seulement aider à trouver l'information, mais également commencer à offrir des moyens pour analyser les données trouvées.

À cet égard, l'une des fonctionnalités les plus intéressantes à développer en complément des outils de recherche habituels concerne l'analyse de l'évolution dans le temps des thèmes qui sont traités dans un corpus. Ce domaine de recherche

semble assez prometteur pour qu'on puisse citer une tentative dans cette direction, celle de la plate-forme ISTEX, qui, à travers le projet ISTEX-R, souhaite développer un outil qui permette, à partir d'un corpus sélectionné par l'usager, de constituer une carte diachronique dont l'objectif final est de permettre à l'usager de mieux caractériser l'évolution des recherches et des connaissances dans le temps. L'objectif est de permettre aux chercheurs de déterminer si l'évolution des recherches et des connaissances dans le temps a connu un bouleversement ou une simple évolution[97]. Il faut signaler toutefois que ce projet, qui intègre l'ensemble des services avancés de la plate-forme ISTEX, est toujours en cours d'élaboration, et qu'il n'est pas possible à l'heure actuelle d'y voir plus que l'espoir d'un futur outil.

DONNER A VOIR L'EVOLUTION DIACHRONIQUE DE L'USAGE DES TERMES DANS UN CORPUS

De façon peut-être moins ambitieuse, on peut citer des établissements qui ont cherché à utiliser sur leurs collections numériques des outils permettant un type d'analyse inspiré par l'application de Google, Ngram Viewer. Précédemment, nous avons vu que les deux principaux concepteurs de l'application, Jean-Baptiste Michel et Erez Aiden, avaient prétendu tirer de cette application une nouvelle science, la « culturomique », non sans susciter le scepticisme de spécialistes en lexicographie. Mais sur un plan technique, les résultats de l'application sont également fragilisés par plusieurs aspects de son corpus de travail, les collections de Google Books. On a ainsi souligné combien les approximations dans la reconnaissance optique de caractères, erronée pour les textes anciens, ainsi que les métadonnées parfois de médiocres qualités, entravaient l'interprétation des résultats de Ngram Viewer[98]. Par ailleurs, la prépondérance numérique des articles scientifiques, nombreux mais peu diffusés, au sein du corpus Google Books, introduit un biais important dans la mesure de l'évolution de la popularité d'un terme[99]. Pourtant, quelle que soit la pertinence de Ngram Viewer appliqué au corpus immense et trop hétérogène de Google Books, l'idée d'appliquer le même mécanisme à des corpus plus réduits, plus homogène, permet d'éviter les difficultés de l'application d'origine. Cette idée a donné lieu à plusieurs initiatives, qui se sont traduites par des exemples de nouveaux instruments de recherche qui offrent des perspectives pour l'analyse et la compréhension des collections. On citera ainsi le travail de Peter Leonard, bibliothécaire de l'université de Yale, qui, en déployant l'outil libre Bookworm sur l'intégralité des collections d'un périodique – les archives du magazine Vogue -, a permis aux chercheurs de faire des recherches sur l'évolution de la fréquence des mots dans ce corpus homogène[100].

[97]On pourra se reporter à la présentation du projet ISTEX-R faite lors du séminaire technique de mars 2015 : Yannick Toussaint (LORIA), Pascal Cuxac (INIST), *ISTEX-R*, mars 2015 (disponible en ligne sur : <http://www.istex.fr/wp-content/uploads/2015/06/9-ISTEX-R-Yannick-TOUSSAINT-et-Pascal-CUXAC.pdf>) (consulté le 8 décembre 2015).

[98]Voir Sarah Shang, « The Pitfalls of Using Google Ngram to Study Language », *Wired*, 12 novembre 2015 (disponible en ligne sur : <http://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/>) (consulté le 5 décembre 2015).

[99]Eitan Adam Pechenick, Christopher M. Danforth, Peter Sheridan Dodds, *Characterizing the Google books corpus : strong limits to inferences of socio-cultural and linguistic evolution*, 7 octobre 2015 (non paginé dans sa version électronique)(disponible en ligne sur : <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0137041>) (consulté le 5 décembre 2015).

[100]Voir cette communication pour l'IFLA 2014 : Peter Leonard, *Mining large datasets for the humanitie*, 2014, p. 5 (disponible en ligne sur : <http://library.ifla.org/930/1/119-leonard-en.pdf>) (consulté le 23 novembre 2015).

Ce que l'on peut attendre de ce type nouveau d'instruments, à mi-chemin entre l'instrument de recherche et celui d'analyse, est de permettre des analyses préparatoires sur des corpus trop abondants. Avec cet outil, l'usager peut ainsi obtenir des courbes représentant les fréquences d'apparition des termes au fil de l'existence du périodique, et comparer, par exemple, la fréquence d'emploi des mots « fille » et « femme » dans le magazine Vogue, pour une recherche dont l'objectif serait d'évaluer l'influence des changements de mentalité sur les textes du magazine[101]. Comme le souligne P. Leonard, les résultats renvoyés par l'outil ne peuvent être considérés comme des éléments probants à eux seuls, mais ils peuvent utilement orienter la recherche par lecture approfondie par la suite. Pour reprendre une terminologie chère aux *digital humanists*, la lecture à distance est ici mobilisée pour préparer une lecture rapprochée et exhaustive.

L'EXEMPLE DE LA MODELISATION DE SUJETS : « LAISSER LES DONNEES S'ORGANISER ELLES-MEMES »

Brève introduction à la « modélisation de sujet »

La recherche d'application de méthodes de fouilles de texte à la gestion des collections peut introduire une certaine rupture par rapport aux pratiques ancrées dans les usages des professionnels de la bibliothéconomie. Ainsi, quand une grande part du traitement documentaire s'applique à inscrire les documents dans un plan de classement pré-établi, ou à indexer le document en utilisant une liste d'autorité pré-existante – c'est à dire qu'il s'agit de rattacher, sauf cas exceptionnel, un nouveau document à ce qui est déjà connu -, ce que laisse entrevoir l'application des méthodes de fouille de texte, c'est, selon l'expression de Peter Leonard, la possibilité de « laisser s'organiser elles-mêmes les données »[102], en dévoilant les motifs et les structures qui les parcourent de façon souterraine. Par cette expression, cet auteur faisait surtout allusion à une famille spécifique de techniques utilisées en TDM, celle des algorithmes de « modélisation de sujet » (*topic modeling*), qui ont connu, et connaissent, une grande popularité parmi les praticiens anglo-saxons des humanités numériques[103]. Ces méthodes d'analyse permettent une première interprétation des thématiques présentes dans les corpus examinés, en générant plusieurs listes de vocabulaires, où se trouvent rassemblés des mots dont la fréquence d'utilisation est corrélée. Disponibles ensuite pour l'analyse humaine, ces nuages de mots permettent à l'observateur d'induire les principaux thèmes traités dans le corpus, et de s'y orienter, l'algorithme pouvant préciser quelle proportion de chaque thème contient un segment déterminé du corpus.

[101] L'exemple est celui de Peter Leonard dans P. Leonard, *Ibidem*, p. 5.

[102] Voir cette communication pour l'IFLA 2014 : Peter Leonard, *Mining large datasets for the humanitie*, 2014, p. 4 (disponible en ligne sur : <http://library.ifla.org/930/1/119-leonard-en.pdf>) (consulté le 23 novembre 2015).

[103] On peut citer, entre autres exemples, les travaux de David J. Newman et Sharon Block, qui ont déterminé, grâce à un algorithme de « modélisation de sujet » les principaux thèmes abordés dans l'ensemble d'une gazette américaine au 18e siècle, et leur évolution diachronique. Voir David J. Newman et Sharon Block, Probabilistic topic decomposition of an eighteenth-century american newspaper (disponible en ligne : http://www.ics.uci.edu/~newman/pubs/JASIST_Newman.pdf). Le travail effectué par Cameron Blevins sur le journal de Marthe Ballard est également un exemple intéressant

De toutes évidences, l'idée d'une « auto-organisation » des « données » telle qu'évoquée par P. Leonard est une hyperbole, et ne doit pas être prise pour argent comptant : ces motifs récurrents que l'analyse informatique dévoile, plutôt que les structures intrinsèques des textes, ne sont que l'approximation qu'en peut atteindre le fonctionnement de l'algorithme, en partant de postulats qui peuvent être contestés. Pas plus qu'une autre technique, les algorithmes de modélisation de sujet ne doivent être maniés sans en connaître le fonctionnement pour en percevoir les limites. Et comme on ne peut guère traiter de ces questions sans entrer dans le détail des programmes informatiques, on comprendra mieux le fonctionnement du processus si on aborde en quelques mots le fonctionnement de l'un d'entre eux, parmi les plus répandus[104] : la modélisation de sujet telle qu'elle s'inspire du modèle probabiliste de l'allocation de Dirichlet latente.

À partir d'un corpus de documents, l'usager commence par indiquer au programme un nombre *n* de thèmes qu'il suppose présent dans les textes. Après avoir enlevé tous les mots vides du corpus (les mots trop communs pour être indexés tels que « le », « la », etc...), l'algorithme opérera ensuite une répartition aléatoire de tous les mots présents dans le corpus en un nombre *n* de liste de vocabulaires, qui représentent déjà, pour le programme, une première approximation anarchique des modèles de sujet présents dans le corpus. Le processus, suivant une démarche itérative, travaillera alors à améliorer cette première répartition des mots : examinant dans son contexte chaque mot de chaque document, il calculera[105] quelle est la probabilité pour qu'il appartienne réellement au modèle de sujet qui lui a été primitivement attribué de façon aléatoire ; si cette probabilité est faible, le programme déplacera le mot à l'intérieur d'une autre liste de vocabulaire, et, ce faisant, aura déjà contribué à améliorer sa première tentative pour établir des listes de modèles de sujets correspondant aux thèmes du texte. Chaque mot étant ainsi examiné et déplacé dans le modèle de sujet qui paraît le plus probable, le programme améliore continuellement les premiers modèles de sujet générés jusqu'à ce que tous les mots soient affectés à la liste de vocabulaires qui paraisse la plus probable pour eux. Le programme, ayant atteint un état stable, présente un ensemble de sacs de mots, composés de termes cooccurrents dans l'ensemble du corpus, où peuvent se lire les différentes thématiques du texte.

Mieux connaître le processus technique et intellectuel derrière ce programme d'analyse sémantique des textes permet d'en percevoir les limites. Ici, une limite paraît évidente : l'algorithme ne peut fonctionner qu'à partir d'une hypothèse,

[104]La popularité de cet algorithme ne fait aucun doute, si l'on en juge par la littérature abondante qui lui est consacrée sur les sites anglo-saxons spécialisés en humanités numériques. On pourra consulter, entre autres, les articles suivants pour une introduction : Scott Weingart, *Topic modeling and network analysis*, 2011 (disponible en ligne : http://www.scottbot.net/HIAL/?p=221) (consulté le 17 octobre 2015) et Matthew L. Jockers, *The LDA Buffet is now open ; or, Latent Dirichlet Allocation for English Major* (disponible en ligne : http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/) (consulté le 23 octobre 2015). Pour des explications consacrées aux opérations statistiques de ces programmes, on consultera plutôt Edwin Chen, *Introduction to Latent Dirichlet Allocation*, 2011 (disponible en ligne : http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/), d'où nous tirons toutes les explications qui suivent sur le fonctionnement des programmes de « modélisation de sujet » comme l'algorithme LDA.

[105]Plus précisément, selon Edwin Chen, *Ibidem*, pour chaque mot *m* d'un document *d*, le programme calcule la proportion de mots du document *d* déjà assignés au thème *t* et la proportion des occurrences du mots *m* assigné dans ce thème *t* pour l'ensemble du corpus. La multiplication de ces deux proportions est le score de probabilité pour le mot *m* appartienne au thème *t*. S'il trouve un autre thème *t* pour lequel ce calcul offre un score plus haut de probabilité d'appartenance pour le mot *m*, alors le mot *m* est déplacé dans cet autre thème *t*. C'est ainsi que le programme prend ainsi en compte les « fréquentations » du mot dans son document – les mots cooccurrents – pour composer ces modèles de sujet. Tel qu'exprimé par Edwin Chen : « And for each topic t, compute two things: 1) p(topic t | document d) = the proportion of words in document d that are currently assigned to topic t, and 2) p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w. Reassign w a new topic, where we choose topic t with probability p(topic t | document d) * p(word w | topic t) (according to our generative model, this is essentially the probability that topic t generated word w, so it makes sense that we resample the current word's topic with this probability). »

hasardeuse, émise par l'utilisateur sur le nombre de modèles de sujet ou de thèmes présents dans le texte. Par ailleurs, il est évident que les listes de mots proposés comme modèles de sujet par le programme nécessitent un observateur pour être interprété, ce qui, parfois, n'est pas facile. Pourtant, les résultats obtenus demeurent intéressants, et à plus d'un titre. Pour reprendre l'exemple de l'analyse du journal de Martha Ballard[106], le programme a bien déterminé une série de sacs de mots qui, aux yeux d'un analyste humain, pouvait tous être interprétés comme des thèmes récurrents du diariste. Ainsi l'un des nuages de mot, contenant les termes « meeting attended afternoon reverend worship », concernait évidemment la vie religieuse de l'écrivaine. Plus encore, le programme permettait, pour cet exemple, d'indiquer dans le corpus où se trouvaient les thèmes traités, et, par la suite, de voir la répartition dans le temps de ces thèmes.

Voici un type de programmes informatiques qui, en gardant présentes à l'esprit les réserves évoquées précédemment, pourraient rendre des services pour présenter une approximation des thèmes présents dans un grand corpus, par exemple de périodiques, au-delà des résultats permis par l'indexation humaine, par ailleurs coûteuse en temps et parfois difficiles à mise en œuvre de façon exhaustive.

Une application intéressante du *topic modeling* : le projet « Mapping text »

En tant qu'instrument de recherche sur des collections de périodique, la modélisation de sujet a fait l'objet de plusieurs expérimentations : David Blei, l'un des chercheurs à l'origine du modèle de l'allocation de Dirichlet latente qui se trouve à la base de nombreux algorithmes de ce type, a par exemple créé une liste de vingt modèles de sujets à partir d'un corpus de l'*American political review*[107]. Nous connaissons également une réalisation très intéressante, le projet américain intitulé « Mapping text »[108]. Pour ce travail, porté conjointement par l'Université du Nord Texas et l'Université de Stanford, l'objectif était d'élaborer une plate-forme pour permettre la recherche et la visualisation sur une collection de périodiques texans. Le corpus numérisé représente 232 567 pages de gazettes dont la publication s'étale de 1829 à 2008.

Pour chaque grande période de l'histoire du corpus, découpé en respectant les scansions de l'histoire texane, dix modèles de sujet de cent mots chacun sont présentés aux usagers, avec leur hiérarchie d'importance. Pour la période couvrant l'ère de la République du Texas, on trouve ainsi des sacs de mots tels que « sales cotton houston received boxes » reflétant l'importance des préoccupations économiques au sein du corpus, ainsi que des listes de type « texas governement state united », où la question politique semble plus prégnante.

Les objectifs du projet n'ont pas tous été atteints, ou quand ils ont été atteints, ses réussites appellent des réserves. Comme le souligne Robert Nelson dans l'article qu'il a consacré au projet[109], les modèles de sujets, pour être tout à fait

[106] Voir Cameron Blevins, *Topic Modelling Martha Ballard's diary*, 1 avril 2010 (Disponible en ligne sur : <http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/>) (consulté le 29 octobre 2015).

[107] Cette expérimentation peut être consultée en ligne à l'adresse suivante : http://topics.cs.princeton.edu/polisci-review/.

[108] Pour une présentation du projet, voir : Andrew J. Torget, Rada Mihalcea, Jon Christencen, Geoff McGhee, *Mapping texts : combining text-mining and geo-visualization to unlock the research potential of historical newspapers* (disponible en ligne sur : <http://mappingtexts.stanford.edu/whitepaper/MappingTexts_WhitePaper.pdf>) (consulté le 23 décembre 2015).

[109] Robert Nelson, « Review of Mapping texts », *Journal of digital humanities*, Vol. 1, N° 3, été 2012 (non paginé dans sa version électronique) (disponible en ligne : <http://journalofdigitalhumanities.org/1-3/review-of-mapping-texts-

profitables, doivent être présentés de façon à servir de porte d'entrée au corpus, afin que les usagers puissent, depuis l'interface où sont présentés les modèles, accéder d'un simple clic aux textes qu'ils décrivent : accéder aux textes permet souvent, d'ailleurs, de mieux comprendre certains modèles de sujet trop opaques. Il s'agit, pour reprendre la terminologie de Moretti, de faciliter le basculement de la lecture distante des collections – par les modèles de sujet – à la lecture minutieuse et exhaustive des textes, la lecture rapprochée. Par ailleurs, le projet est également entravé par la qualité des textes océrisés d'origine, qui, là encore, sur un corpus immense – tous les périodiques texans sur deux siècles - constituent un véritable obstacle.

Toutefois, il reste que ce projet jette des perspectives intéressantes sur la façon dont des techniques de TDM pourraient permettre la constitution d'instruments de recherches et d'analyses complémentaires aux dispositifs déjà en place. La modélisation de sujets, quand les modèles sont suffisamment clairs – à l'exemple de ceux obtenus par D. Blei sur la collection de l'*American political review*[110] - permettent une première approche des thématiques présentes dans un corpus, et surtout permettent de localiser à quel endroit du corpus une thématique précise se trouve abordée.

by-robert-nelson/#review-of-mapping-texts-by-robert-nelson-n-1>) (consulté le 22 octobre 2015). « It is always useful and sometimes essential to be able to move from the word distribution of a topic to the documents that are highly representative of that topic to get a more subtle sense of its substance and meaning, a point Lisa Rhody thoughtfully illustrates in a recent blog post. »

[110]Par exemple, voici deux des modèles de sujet obtenus par David Blei sur l'*American political review* : « voting voters vote voter candidate » et « states war international military soviet ». Voir la liste des modèles de sujet à l'adresse suivante : http://topics.cs.princeton.edu/polisci-review/.

DES OUTILS POUR LE TRAITEMENT DOCUMENTAIRE

La croissance des collections numériques implique naturellement la mise au point de méthodes de traitement adaptées à ces types de collection. La fouille de texte, entendue au sens le plus large, peut fournir de façon évidente une assistance pour administrer ces fonds documentaires, quoique, dans une certaine mesure, nous verrons que les techniques automatisées représentent surtout, dans l'état actuel de l'art, une méthode pour accomplir un travail préparatoire, et non une façon de remplacer l'expertise humaine.

Il reste que sur des tâches réputées fortement consommatrices en temps, comme l'indexation de documents ou l'alimentation de référentiels terminologiques, l'usage des acquis de la fouille de texte peut rendre des services. Comme ces procédures ne sont toutefois pas encore très développées, nous nous contenterons de présenter deux types de tâches de fouille appliqués aux collections numériques, avec des exemples d'application de ces tâches, sans viser l'exhaustivité.

LA CLASSIFICATION AUTOMATIQUE DES TEXTES

Le classement automatique des textes est l'une des tâches les plus classiques de la fouille de texte, utilisée à la fois pour des applications commerciales et dans des contextes d'études en sciences humaines. Ce type d'application du TDM présente un intérêt évident pour assister les chargés de ressources numériques : sous sa forme dite supervisée (c'est à dire quand le chercheur dispose d'un corpus de départ sur lequel entraîner l'algorithme, qui ensuite cherchera à identifier dans un ensemble plus vastes les documents présentant des caractéristiques semblables), le classement automatique de texte peut servir à retrouver dans un ensemble les documents répondants à ces caractéristiques précises.

Nous connaissons au moins deux exemples de ce type d'usage, qui tous deux interviennent dans des contextes où la forte croissance documentaire incite l'application de ce type de méthodes de traitement massif. À l'Institut national de recherche agronomique (INRA), la classification automatique de textes a été employée pour répondre à un besoin d'évaluation bibliométriques : il s'agissait de repérer les articles dont la publication s'inscrivait dans le cadre d'un programme de recherche, bien que ce programme ne soit pas cité dans le texte de l'article[111]. Un petit corpus de trente-six publications, repérées comme issues du programme de recherche dont l'activité était évaluée, a donc été utilisé pour calculer l'empreinte sémantique des ressources à trouver, et à permis, par la suite, de repérer 793 articles susceptibles d'appartenir à l'ensemble de ce programme de recherche. Il s'agissait, ici, d'une expérimentation, dont l'un des acteurs principaux a souligné qu'elle devrait être poursuivie avec la collaboration de chercheurs pour valider le corpus noyau et le corpus élargi, mais dont les premiers résultats semblent intéressants.

Par ailleurs, un projet mené sur les fonds numériques de la bibliothèque Hathi Trust suggère une autre utilisation possible des algorithmes de classification.

[111]Toutes les informations présentées ici sur le TDM à l'INRA ont été obtenues au cours d'un entretien que M. Mathieu Andro, ingénieur à l'INRA, nous a aimablement accordé. Plus précisément, il s'agissait de repérer les articles publiés dans le cadre du métaprogramme ACCAF (Adaptation de l'agriculture et de la forêt au changement climatique).

La bibliothèque était confrontée au problème suivant : une absence de métadonnées concernant les dates de nombreux documents numériques au sein de ses fonds. À partir de ce constat, un groupe de chercheurs a établi avec succès un algorithme de classement pour prédire approximativement la date de publication des textes dont les métadonnées manquaient [112]. Pour ce faire, l'algorithme s'appuyait sur plusieurs caractéristiques des textes à classer dans le corpus, dont les erreurs d'OCR – certaines erreurs étant caractéristiques d'une période de publication -, ainsi que les caractéristiques lexicales. Selon les membres de l'équipe de recherche, les résultats obtenus sont satisfaisants, tout en gardant une marge de progression[113]. Bien que nous ne puissions pas en juger, force est de constater qu'il s'agit d'un exemple intéressant d'application d'algorithmes de classifications sur les textes numériques pour améliorer l'administration d'un corpus numérique.

LES APPLICATIONS DE L'EXTRACTION D'ENTITES NOMMEES : INDEXATION AUTOMATIQUE ET EXTRACTION TERMINOLOGIQUE

La reconnaissance d'entités nommées est l'une des tâches de fouille de texte les plus utiles pour assister les professionnels de la documentation dans l'enrichissement des métadonnées liées aux documents. La reconnaissance d'entités nommées peut entrer dans processus différents : l'indexation automatique, et l'enrichissement de référentiels terminologiques.

L'alimentation semi-automatisée de thésaurus est une autre application potentielle de la reconnaissance d'entités nommées, où des outils de traitement automatisé du langage accomplissent dans le corpus un travail de repérage des « candidats-termes », c'est à dire de termes qui, après vérification d'un spécialiste, pourront intégrer un référentiel, que celui-ci soit un thésaurus ou une ontologie[114]. Dans ce type d'application, les textes subissent d'abord une première étape de traitement afin de permettre l'analyse : séparation du texte en phrases, puis en unités plus simples, lemmatisation, et enfin étiquetage morpho-syntaxique. C'est à partir de cette dernière étape que l'extracteur va travailler, en appliquant au texte ainsi étiqueté des règles de détection fondées sur les catégories morpho-syntaxiques : il relèvera ainsi, par exemple, l'expression « protéine de soja », parce que correspondant à un motif, une séquence Nom-Préposition-Nom qu'on lui a demandé de relever. À l'issue de ce travail, un analyste humain devra prendre en charge la liste des éléments repérés par l'extracteur, pour déterminer quels sont les plus pertinents pour alimenter le thésaurus.

Les acteurs de la plate-forme ISTEX travaillent actuellement à la mise en place d'un tel extracteur de termes. Bien que l'extraction terminologique ait fréquemment pour objectif l'alimentation d'un thésaurus, le projet ISTEX vise un

[112]Ce travail est présenté dans l'article suivant : Siyuan Guyo, Trevor Edelblute, Bin dai, Miao Chen, Xiaozhong Liu, « Toward Enhanced Metadata Quality of Large-Scale Digital Libraries: Estimating Volume Time Range », *iConference 2015 Proceedings*, 2015 (disponible en ligne : https://www.ideals.illinois.edu/bitstream/handle/2142/73656/186_ready.pdf?sequence=2) (consulté le 20 décembre 2015).

[113]Voir Siyuan Guyo, Trevor Edelblute, Bin dai, Miao Chen, Xiaozhong Liu, *Ibidem*, p. 10 pour les grilles de résultats.

[114]Pour les informations présentées dans ce paragraphe, nous sommes gré à M. Gaël Guibon, membre de l'INIST-CNRS, chargé du projet d'extraction terminologique de la plate-forme ISTEX, qui a bien voulu répondre à nos questions sur son travail.

autre objectif : le repérage des termes les plus pertinents pour un domaine ainsi que leur emplacement.

L'indexation automatique est une autre des applications de la reconnaissance d'entités nommés. Dans l'exemple de la plate-forme ISTEX, l'indexation automatique concerne plusieurs types d'entités : les noms de personnes, de lieux, d'organisation et les dates. Ces types d'entités sont souvent reconnaissables grâce au contexte local, c'est à dire les mots qui les précèdent ou leur succèdent[115]. Pour le projet ISTEX, cette reconnaissance des entités nommées permet un balisage automatique des textes en XML/TEI, qui permet par la suite une recherche plus puissante sur les corpus.

Qu'il s'agisse de l'indexation automatisée des textes ou de l'extraction terminologique, il faut souligner que ces processus interviennent surtout en tant qu'assistance informatique à l'analyse humaine. Les candidats-termes proposés par le programme doivent être vérifiés par un expert pour se voir attribuer le statut de vedette-matière, et les termes balisés automatiquement par le système sont vérifiés.

[115] Voir compte-rendu du séminaire technique ISTEX, page 12, disponible en ligne à l'adresse suivante : <http://www.istex.fr/wp-content/uploads/2015/06/CR-s%C3%A9minaire-services-ISTEX-18-et-19-mars-2015_V080615.pdf>.

LE POIDS DES INCERTITUDES JURIDIQUES SUR LE TEXT MINING

Qu'on la voit modestement comme la version automatisée d'une activité manuelle de relevé d'informations[116], ou de façon plus ambitieuse comme une nouvelle méthode de recherche à part entière, qui permettrait de repérer des corrélations ténues restées jusqu'alors inaperçues entre phénomènes dans la littérature, la fouille de texte et de données ouvre de nombreuses perspectives à la recherche scientifique. Cependant, la situation de ces techniques d'analyse aujourd'hui en Europe peut sembler à certains égards paradoxale. Bien que cette technique soit créditée de nombreuses vertus[117], et soutenue par l'Union européenne dans le financement qu'elle accorde aux deux projets FutureTDM et Open Minted[118], le constat s'impose cependant d'une faiblesse relative de son développement. Le consortium FutureTDM identifie les obstacles suivants à sa diffusion[119] : du côté des communautés scientifiques, une certaine absence de prise de conscience de l'intérêt de cette technologie pour la recherche, et, quand ce n'est pas le cas, des incertitudes juridiques qui imposent des efforts disproportionnés pour sécuriser un projet de fouille ; un manque de compétences par rapport aux besoins, qui traverse à la fois le domaine de la recherche et le monde des bibliothèques ; et enfin, *last but not least*, le fait qu'en l'état du droit de la propriété intellectuelle européen, la fouille peut constituer une atteinte aux droits des ayant-droits. Le site finit en notant également que le système de licence proposé par les éditeurs soulève des interrogations.

Nous avons déjà par ailleurs abordé les difficultés d'ordre technique, qui peuvent entraver le développement du text mining. Il faut désormais s'attacher à décrire les problèmes d'ordre juridique, qui représentent pour le text mining en Europe le principal frein à l'extension de sa pratique. L'absence de clarification sur la légalité du TDM quand il porte sur des documents encore protégés par le droit d'auteur amoindrit de façon évidente l'étendue des recherches possibles, puisqu'elles doivent alors se limiter aux documents du domaine public.

Avant de revenir plus en détails sur les étapes du débat et son actualité, synthétisons-le. La controverse pourrait se résumer à une divergence sur le statut qu'il convient d'accorder au text mining : une partie des ayants-droits, parmi lesquels se sont surtout faites remarquer les grandes maisons d'édition scientifiques, estime que la fouille de données représente un nouveau type d'exploitation des contenus sous droit, qui n'est pas naturellement inclus dans le droit d'accès aux documents que les usagers et leurs représentants négocient quand ils font l'acquisition des ressources. La solution proposée par ce type d'acteurs est

[116]Pierre-Carl Langlais, « Data Mining : quand Elsevier écrit sa propre loi », *Sciences communes*, 8 février 2014 (Disponible en ligne sur : <http://scoms.hypotheses.org/98>) (consulté le 6 novembre 2015). « Nous faisons, tous, au quotidien du data-mining en récupérant des informations dans des textes préalablement publiés. L'API et les algorithmes d'extraction ne sont que des outils supplémentaires visant à simplifier cette activité fondamentale. Un projet de l'ampleur de Text2Genome aurait très bien pu être réalisé avec un papier et un crayon : son élaboration aurait simplement pris un ou deux siècles plutôt que quelques années. »

[117]Voir Ligue des bibliothèques européennes de recherche, *The perfect swell : defining the ideal conditions for the growth of text and data mining in Europe*, 2013, (disponible en ligne sur: <http://libereurope.eu/wp-content/uploads/TDM%20Workshop%20Report%5B1%5D.pdf>) (consulté le 10 décembre 2015). Page 2 : « Pour le dire simplement, la fouille de texte sauve des vies ».

[118]FutureTDM et OpenMinted sont deux projets soutenus par l'Union européenne pour promouvoir la fouille de texte, entre lesquels les tâches sont ainsi réparties : FutureTDM, consortium rassemblant plusieurs acteurs de la recherche européenne dont LIBER, doit identifier les freins au développement du TDM en Europe puis proposer des recommandations pour les dépasser. OpenMinted se propose de construire une infrastructure pour la découverte et l'utilisation des technonologies de texte et data mining.

[119]Voir la partie « Overview » du site suivant : http://project.futuretdm.eu/project-overview/

donc d'établir une voie contractuelle pour la fouille de texte, qui passe par la souscription d'une offre spécifique par les usagers qui souhaitent appliquer ces méthodes d'analyse aux corpus. Les opposants à la « voie contractuelle », parmi lesquels des représentants des associations professionnelles de bibliothécaires ainsi que des collectifs de scientifiques[120], font valoir de leurs côtés que la fouille de contenu ne diffère pas essentiellement du relevé manuel d'informations, et que « le droit de lire devrait entraîner le droit d'extraire »[121]. Ce dernier groupe milite donc pour que la fouille de texte soit considérée comme une nouvelle exception au droit d'auteur[122].

Tant que le législateur n'a pas statué sur le statut sur la fouille de texte ou de données, la position des éditeurs prévaut naturellement, puisqu'ils contrôlent l'accès aux textes. Les situations sont néanmoins contrastées selon les pays et les traditions juridiques, et nous en ferons le tour. Ainsi, les États-Unis ont reconnu que la fouille de texte relevait des exceptions au droit d'auteur inclus dans la catégorie de l'usage raisonnable (« fair use »).

Cette question de la légalité des offres de text and data mining est loin d'avoir trouvé sa résolution, du moins en Europe. Nos ambitions se bornent ici à chercher à offrir une synthèse à jour des débats en cours.

L'offre des éditeurs en matière de text mining

Pour comprendre la situation actuelle, il faut revenir à la décision inaugurale de l'éditeur Elsevier, qui le premier en janvier 2014 rendit public des conditions spécifiques pour effectuer des opérations de fouille de textes sur ses collections de périodiques électroniques[123]. Cette décision fut accueillie par une partie de la communauté scientifique comme une avancée ; elle rompait avec une situation antérieure marquée par l'absence de cadre formalisé, qui se traduisait par des décisions prises au cas par cas par les éditeurs, et des procédures d'autorisations longues[124]. C'est à cette situation que les animateurs du projet Text2Genom, Max Hauessler et Casey Bergman, attribuaient le temps pris pour obtenir les autorisations de fouiller les articles dont ils avaient besoin pour leur projet[125]. C'est pourquoi la prise de position d'Elsevier semblait promettre un développement de la pratique du TDM, en allégeant la procédure.

Cependant, les conditions que l'éditeur Elsevier fixait pour fouiller ses corpus électroniques valaient prise de position sur le statut juridique de cette technique, et constituaient un précédent que certains jugèrent immédiatement

[120]Parmi les prises de position les plus notables contre la voie contractuelle, on peut souligner celle de la Ligue des bibliothèques européennes de recherche (LIBER), ainsi que celle de l'Association des directeurs de bibliothèques universitaires (ADBU). Le collectif SavoirsCom1, engagé dans la protection des, est un autre exemple de structures s'étant prononcées contre la politique des licences.

[121]« The right to read should be the right to mine. » Voir Ligue des bibliothèques européennes de recherche, *Ibidem*, p. 3.

[122]C'est la position à la fois défendue par Liber et SavoirsComm1.

[123]Voir Chris Shillum, *Elsevier updates text-mining policy to improve access for researchers*, 31 janvier 2014, (disponible en ligne sur : < https://www.elsevier.com/connect/elsevier-updates-text-mining-policy-to-improve-access-for-researchers>) (consulté le 2 décembre 2015).

[124]Voir Richard van Noorden, « Elsevier opens its paper to text-mining », *Nature*, 3 février 2014, (disponible en ligne sur : <http://www.nature.com/news/elsevier-opens-its-papers-to-text-mining-1.14659>) (consulté le 2 décembre 2015).

[125]Voir Richard van Noorden, « Trouble at the text mine », *Nature*, 7 mars 2012 (disponible en ligne sur : <http://www.nature.com/news/trouble-at-the-text-mine-1.10184>) (consulté le 2 décembre 2015).

fâcheux[126]. Le texte détaillant cette nouvelle politique[127] précise en effet que le droit de TDM intégrait la licence standard d'abonnement à la base ScienceDirect, et devait être spécifiquement incluse dans les licences renouvelées. Cette formulation constituait donc très clairement la fouille de texte comme droit distinct de celui d'accéder à un contenu et de le lire, devant, en tant que tel, faire l'objet d'une concession séparée des droits d'accès à la littérature électronique. De sérieuses limitations techniques sont par ailleurs imposées : les chercheurs appartenant à une institution abonnée à ce service doivent, pour en profiter, passer exclusivement par l'interface de programmation (ou API) fournie[128] par l'éditeur, ce qui n'est possible qu'en s'inscrivant au préalable auprès de ses services, procédure qui impose pour le chercheur de donner des détails sur son projet de recherche. Cet état de fait semble conférer un pouvoir décisionnel des éditeurs sur les projets de recherche, et fait planer la menace d'une marchandisation ultérieure des données ainsi récoltées[129]. D'autres restrictions sont prévues : ainsi, les citations extraites des corpus par fouille de texte ne doivent pas dépasser les 200 caractères, et cette disposition, introduite vraisemblablement pour rendre impossible la reconstitution des articles à partir des éléments prélevés par fouille, va à l'encontre du droit de courte citation[130]. Aux termes de la licence, les auteurs des articles exposant des éléments extraits par text mining doivent également publier ces éléments en les plaçant sous une licence non commerciale de type Creative Commons, et donner l'origine des données extraites en signalant le DOI[131] de chaque donnée. L'exigence d'une licence non-commerciale pour la publication des données obtenues par la fouille de texte n'est pas sans provoquer des interrogations sur ses conséquences : par ce biais, l'éditeur semble vouloir se faire reconnaître un droit de propriété sur les données, et contrôler leurs réutilisations, alors qu'un principe fondamentale du droit de la propriété intellectuelle est qu'il couvre l'expression originale d'une œuvre, non les informations qu'elle peut contenir[132].

On voit que la proposition d'Elsevier, parce qu'elle contient de façon implicite une prise de position sur le statut juridique de la fouille de données comme usage distinct du droit de lecture, a des implications qui dépasse la simple question de l'accès aux collections électroniques de cet éditeur. Celui-ci a d'ailleurs

[126]Voir Richard van Noorden, *Ibidem*, 3 février 2014 « But some researchers feel that a dangerous precedent is being set. They argue that publishers wrongly characterize text-mining as an activity that requires extra rights to be granted by licence from a copyright holder, and they feel that computational reading should require no more permission than human reading. »

[127]Voir Chris Shillum, *Elsevier updates text-mining policy to improve access for researchers*, 31 janvier 2014, (disponible en ligne sur : <https://www.elsevier.com/connect/elsevier-updates-text-mining-policy-to-improve-access-for-researchers>) (consulté le 2 décembre 2015).

[128]Une interface de programmation se dit en anglais Application programming interface, d'où l'acronyme API largement utilisé désormais dans la littérature française. Une interface de programmation a pour fonction d'offrir un intermédiaire entre différents programmes ou services, ou encore de servir « de façade par laquelle un logiciel offre des services à un autre logiciel » (Article interface de programmation de Wikipedia). Dans le cas présent, l'API sert à accéder aux données contenus dans un réservoir de données, ici les collections électroniques de Wikipedia. En conformité avec l'usage, nous utiliserons désormais le terme d'API.

[129]Sur cette question, voir Grégory Colcanap, Christophe Perales, *CSPLA – Mission relative au data mining (exploration de données)*, ADBU, 2014, p. 4 « Cela revient à donner aux éditeurs de contenus, qui ne sont pour rien dans le financement des projets de recherche, le droit de décider quelle recherche pourra ou non voir le jour. ».

[130]Pierre-Carl Langlais, « Data Mining : quand Elsevier écrit sa propre loi », *Sciences communes*, 8 février 2014 (Disponible en ligne sur : <http://scoms.hypotheses.org/98>) (consulté le 6 novembre 2015). « La première condition contredit ouvertement l'exception de courte citation. En France, le législateur a volontairement adopté une définition floue afin de laisser le juge statuer sur chaque cas selon ses spécificités. (…) Or Elsevier tente de contourner le droit existant pour imposer sa propre règle : une limite de 200 caractères, purement arbitraire. »

[131]Le DOI est le Digital object identifier, « un mécanisme d'identification de ressources, qui peuvent être des ressources numériques, comme un film, un rapport, des articles scientifiques, mais également des personnes ou tout autre type d'objets. » Source : article DOI de Wikipedia

[132]Pierre-Carl Langlais, *Ibidem*, 8 février 2014. « Elsevier impose discrètement l'idée que les informations pourraient être protégées. [...] Ni les informations, ni les données « solitaires » n'échappent à cette appropriation globale : toute reprise du contenu d'Elsevier à des fins de data mining devra être publié sous une licence non-commerciale. »

fait école, puisque, par la suite, un éditeur comme Springer a également demandé à ce qu'un formulaire soit rempli, et des informations sur les projets de recherche données, pour pouvoir procéder au TDM sur ses collections[133]. Sur le plan théorique, on a aussi pu faire remarquer que si la fouille de texte devait être un usage accordé par les ayant-droits, la fouille du Web serait chose impossible dans un cadre légal, en raison de la multiplicité des intervenants à contacter par les chercheurs[134]. Envisagée sous l'angle de ses conséquences techniques, la voie contractuelle proposée au text mining par l'octroi de licence, si elle se généralisait, ne serait donc pas sans poser problème. Elle suppose que pour des projets de fouilles de textes impliquant l'analyse de documents issus des collections de plusieurs éditeurs, l'équipe de recherche obtienne de tous les éditeurs le même droit. Cela implique donc un travail préliminaire important pour obtenir de chacun des éditeurs ce droit, et impose avant le même le travail de recherche proprement dit une importante dépense d'énergie consacrée à sa préparation administrative[135]. Par ailleurs, des projets de fouille à grande échelle nécessitent d'opérer simultanément sur les différents corpus, et non d'interroger successivement chacune des collections détenues par des éditeurs différents. L'octroi séparé de licences risque d'entériner la répartition arbitraire des articles entre éditeurs comme des secteurs de la recherche.

Il est vrai qu'en réponse à ces difficultés, un projet d'API mutualisée entre éditeurs a vu le jour : il s'agit du projet Prospect, initié par CrossRef[136], projet devenu depuis les « Text and data mining services » de Crossref. Telle que présentée par ses concepteurs[137], cette interface de programmation se propose de résoudre une partie des difficultés que pose la fouille de texte sur des collections issues d'éditeurs différents : éviter la négociation bilatérale des licences de part et d'autre, et permettre la fouille simultanément sur des revues de plusieurs éditeurs distincts, ainsi que sur des revues open-access, en offrant une API unique pour l'ensemble de ces collections. Le service a été lancé en mai 2014, et l'absence d'avis porté sur son fonctionnement[138] empêche pour l'instant de dire si ce service dément le scepticisme dont faisaient preuve certains acteurs concernant l'hypothèse d'une API mutualisée entre grands éditeurs [139].

Quoi qu'il en soit de la réussite ou de l'échec de cette API mutualisée, il est évident qu'elle ne peut résoudre les questions préjudicielles sur le statut légal de la fouille de texte[140]. La position des éditeurs, qui veut que la fouille de données soit

[133]Voir Gregory Colcanap, Christope Perales, *Ibidem*, p. 5.

[134] Cet argument avancé contre l'octroi du droit de fouille par licence se trouve dans Grégory Colcanap, Christophe Perales, *CSPLA – Mission relative au data mining (exploration de données) : l'analyse de Couperin et de l'ADBU*, p.3 (disponible en ligne : http://adbu.fr/wp-content/uploads/2014/04/Audition_CSPLA_TDM_2014_04_04_final.pdf) (consulté le 16 décembre 2015), ainsi que dans Pierre-Carl Langlais et Lionel Maurel, *Quel statut légal pour le content-mining ?*, p. 15 (disponible en ligne sur : http://www.savoirscom1.info/wp-content/uploads/2014/01/Synthe%CC%80se-sur-le-statut-le%CC%81gal-du-content-mining.pdf).

[135]Gregory Colcanap, Christophe Perales, *Ibidem*, p. 5

[136]CrossRef est un organisme à but non lucratif dont le but historique est d'assurer la gestion des identifiants d'objets numriques (DOI, Digital Object Identifier), en garantissant la liaison entre un identifiant et le texte intégral d'un article. Son rôle est donc d'assurer la pérennité des mécanismes de citation, en effectuant la redirection entre la citation d'un texte par son DOI dans un article, et la localisation actuelle d'un article plein texte, sujette à varier au gré des changements de serveur. On compte parmi ses participants de nombreux acteurs de l'édition scientifique ainsi que des bibliothèques universitaires. Voir la présentation de l'organisme à l'adresse suivante : http://www.crossref.org/01company/02history.html

[137]Voir la section de présentation la page de présentation du projet : http://tdmsupport.crossref.org/.

[138] Nous n'avons malheureusement pas pu trouver d'évaluation du fonctionnement de ce service pendant la rédaction de ce travail.

[139]Gregory Colcanap, Christophe Perales, *Ibidem*, p. 3 : « Il n'existe pas de licence globale ou de proposition d'une API mutualisée entre les acteurs de l'édition et du Web, et il n'y en aura pas avant longtemps».

[140]Pas plus, d'ailleurs, que la réussite d'une telle API ne permettrait de fouiller le Web légalement, s'il était reconnu que le droit de fouille devait être concédé par les ayant-droits pour être légal.

une activité fondamentalement différente du relevé manuel d'information, représente le premier point de divergence apparu entre bibliothécaires et éditeurs : la fouille de contenus, selon le compte-rendu d'un atelier organisé par LIBER, la ligue européenne des bibliothèques publiques, est tout simplement « une forme de lecture », et « le droit de lire devrait entraîner le droit d'extraire »[141]. À la suite de cette première déclaration, Pierre-Carl Langlais et Lionel Maurel ont également souligné une continuité fondamentale entre la fouille de données et les pratiques habituelles des chercheurs, qui font depuis toujours de la fouille de contenu, « avec un crayon et un papier »[142]. Les mêmes auteurs soulignent que ce qui, dans la fouille de texte, fait l'objet d'une extraction, n'est pas éligible à la protection du droit de la propriété intellectuelle : ce dernier, toutes traditions juridiques confondues, cherche avant tout à protéger l'expression originale des idées, non l'information ou la donnée qui s'y trouve enchâssée, celle que la fouille de texte a précisément pour but de récupérer[143]. En mai 2015, ces prises de position, jusqu'alors déterminées mais éparses, ont fini par trouver une expression plus structurée dans la déclaration de la Haye sur l'extraction de connaissances à l'ère numérique[144], dont les articles et propositions concernent la fouille de contenus : s'y trouvent notamment affirmés la distinction entre les données et les textes, ainsi que le principe que le « droit de lire est le droit d'extraire ».

Cette contestation du point de vue des éditeurs n'est pas sans reconnaître cependant que le text mining pose des problèmes juridiques, sinon au regard de ses objectifs, du moins à celui de ses conditions techniques de possibilité. En effet, une tâche de text mining peut nécessiter que l'intégralité du corpus fouillé fasse l'objet d'une copie : s'agissant de très gros volumes de données, l'analyse algorithmique ne peut opérer sur des serveurs distants. C'est cette nécessité d'une copie transitoire qui, le cas échéant, fait tomber la fouille sous le coup des dispositions relatives au droit d'auteur, et surtout suscite les réticences des grands éditeurs scientifiques face à ces processus, qui leur font craindre des opérations de duplication massive de leurs ressources bases de données. Neutre dans ses objectifs, la fouille peut donc potentiellement représenter durant les opérations une « illégalité collatérale »[145], selon l'expression de Pierre-Carl Langlais et Lionel Maurel.

Des situations contrastées selon les traditions juridiques

Légale quant à ses buts, mais pouvant nécessiter pour s'accomplir une copie des textes jugée délictueuse, il est certain que la fouille de texte pose au droit de la

[141]Voir Voir Ligue des bibliothèques européennes de recherche, *Ibidem*, p. 3.

[142]Pierre-Carl Langlais et Lionel Maurel, *Quel statut légal pour le content-mining ?*, p. 9 (disponible en ligne sur : http://www.savoirscom1.info/wp-content/uploads/2014/01/Synthe%CC%80se-sur-le-statut-le%CC%81gal-du-content-mining.pdf). Voir également page 9 : « Si le content-mining marque un changement d'échelle, il ne fonde pas une activité nouvelle. Extraire et synthétiser des informations préexistantes constituent le labeur quotidien du chercheur depuis que la recherche scientifique existe. ».

[143]Voir Pierre-Carl Langlais et Lionel Maurel, *Ibidem*, p. 9 : « Par définition, le droit d'auteur ou le copyright ne portent pas sur des informations « brutes », mais sur leur expression originale. Ce principe se retrouve dans toutes les législations. » Ce problème de la justification de l'invocation du droit de la propriété intellectuelle pour empêcher le text mining était déjà souligné par les conclusions de l'atelier de recherche sur le text mining de la La British Library :« Logically text and data mining should not fall under the scope of copyright law as it is not concerned with the artistic expression of ideas which copyright regulates, but with the extraction and analysis of facts and data which are specifically excluded from regulation by IPRs in international laws. » (voir Ligue des bibliothèques européennes de recherche, *Ibidem*, p. 3).

[144]Le texte de la Déclaration se trouve en ligne à l'adresse suivante : http://thehaguedeclaration.com/

[145]Voir Pierre-Carl Langlais et Lionel Maurel, *Ibidem*, p. 12.

propriété intellectuelle un problème, comme toute innovation technologique introduisant des usages qu'un état antérieur du droit n'avait pas prévu. Les réponses à ce problème ont varié selon les contextes nationaux et les traditions juridiques : en complément de l'analyse des questions théoriques posées par cette technique, passons désormais à un rapide examen de la situation légale de la fouille de texte dans le monde. Concernant les réponses légales au text mining, on a pu établir la typologie suivante, comprenant quatre catégories[146] : les pays qui s'inscrivent dans une tradition du droit d'auteur français, où le text mining n'est pas autorisé (catégorie qui inclut la plupart des pays européens, dont la France) ; les pays dont le droit de la propriété intellectuelle s'inspire du principe du « copyright », dans sa version britannique, pour qui le text mining n'est « probablement pas autorisé » ; les pays qui s'inspirent de la version américaine du copyright, pour qui le text mining est « probablement autorisé » ; le quatrième groupe étant celui des pays où le text-mining est autorisé avec certitude, car le législateur a introduit une exception légale dans la loi, cette dernière catégorie incluant le Japon et la Grande-Bretagne.

L'importance dans les débats autour du TDM des concepts de « fair use » et de « fair dealing » nous incite à nous concentrer dans un premier temps sur les législations anglaises et américaines, qui toutes deux ont en commun d'autoriser le TDM, avant d'évoquer dans un second temps le cas de l'Union européenne, où le TDM reste illégal en dehors du cadre contractuel.

Les pays où le text mining est autorisé : « fair dealing » et « fair use »

La Grande-Bretagne a inclus le text mining parmi les exceptions au droit d'auteur par des mesures entrées en vigueur le 1er octobre 2014. Le texte de la loi[147] stipule désormais qu'un chercheur, s'il bénéficie d'un accès légal à des textes, peut également y effectuer des opérations de fouille de texte dans le cadre de recherches sans but commercial. Il est également autorisé à copier de grandes quantités de données sous droit. Le texte précise néanmoins que cette pratique doit rester dans le cadre de ce que la tradition juridique britannique appelle un usage équitable (*fair dealing*) : c'est à dire qu'elle ne doit pas avoir pour conséquence d'entraîner des pertes financières pour l'ayant-droit, et doit être proportionnée au but de la recherche.

On notera que la mise en place de cette exception, si elle a accordé une sécurité juridique aux chercheurs, n'a pas introduit de bouleversement majeur dans les pratiques des éditeurs : ainsi, même si légalement les chercheurs britanniques ne sont plus dans l'obligation de faire une demande spécifique à Elsevier pour fouiller les corpus de Science Direct – un accès légal au texte impliquant le droit d'analyse algorithmique - Elsevier continue à imposer l'usage de son API pour télécharger son contenu. En effet, la loi britannique stipule que les ayant-droits peuvent introduire des restrictions d'accès pour ne pas surcharger les serveurs, ce

[146]Christian Handke, Lucie Guibault, Joan-Josep Vallbé, *Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research*, 7 juin 2015 (disponible en ligne sur : <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2608513>) (consulté le 3 décembre 2015).

[147]On trouvera un résumé des nouvelles exceptions au droit d'auteur en faveur du text mining dans cette brochure du gouvernement britannique : Intellectual property office, *Exceptions to copyright : research*, octobre 2014 *(*Disponible en ligne à cette adresse : https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf) (Consulté le 3 décembre 2015).

qui donne à l'éditeur le pouvoir d'imposer l'usage de son API[148], présentée comme un moyen de réguler le trafic.

Le text mining est également légal aux Etats-Unis, depuis qu'il a été reconnu comme l'un des usages d'œuvres protégées qui intègre le domaine d'un « usage raisonnable »[149]. Cet état de fait découle de la jurisprudence rendue dans l'affaire qui opposait Google Books à l'Authors Guild, la société américaine de défense des droits d'auteur, dans une procédure qui initialement contestait la démarche de numérisation de Google, avant que les débats ne s'élargissent à la légitimité des procédures de fouille de texte[150]. Cette procédure, qui a commencé en 2005, a connu son dernier épisode en octobre 2015, lorsqu'une cour d'appel a confirmé les décisions des précédents jugements. Il est intéressant de noter que pour justifier sa décision d'inclure la fouille parmi les usages équitables, le juge a fait valoir que cette pratique représentait un usage « transformatif », soit un travail sur le contenu qui implique un apport original.

Cette décision est de grande conséquence : la fouille de texte est donc légitime sur des contenus protégés par droits d'auteurs pour en extraire de la donnée.

L'Union européenne

Les cas britanniques et américains ont été évoqués en priorité, car ces deux cadres législatifs pour le TDM font, à des degré divers, figure de référence. La crainte d'un retard pris par la recherche européenne par rapport à ces deux pays est par ailleurs fréquemment exprimée, et représente certainement l'une des raisons pour qu'une évolution de la législation européenne sur le TDM soit annoncée en 2016[151]. En effet, cette dernière est bien moins accueillante à cette nouvelle pratique de recherche, et le TDM s'y trouve encore dans une situation d'incertitude juridique qui explique un retard constaté par certains auteurs dans les usages des chercheurs européens[152].

Il y a à cet état de fait des raisons qui concernent profondément les traditions juridiques des pays concernés, et notamment le fait que le droit de la propriété intellectuelle, qui inspire la législation européenne, définisse traditionnellement de façon plus restrictive les exceptions au droit d'auteur que son équivalent en pays anglo-saxons, le copyright. Comme le notent les auteurs du rapport sur le TDM remis à la commission européenne en 2014, la tradition européenne du droit d'auteur ne souffre qu'un nombre limité d'exceptions précisément décrites, quand le copyright se contente de fournir au juge un concept souple, celui d'un « usage équitable », pour définir si un usage intègre ou pas le

[148]Voir la présentation de sa politique sur le site de l'éditeur : Gemma Hersh, *How does Elsevier's text mining policy work with new UK TDM law ?*, 9 juin 2014 (disponible en ligne sur : <https://www.elsevier.com/connect/how-does-elseviers-text-mining-policy-work-with-new-uk-tdm-law>) (consulté le 5 décembre 2015).

[149]Le « fair use », que l'on traduit en France par « usage loyal », ou « usage raisonnable », est un « ensemble de règles de droit, d'origine législative et jurisprudentielle, qui apportent des limitations et des exceptions aux droits exclusifs de l'auteur sur son oeuvre » (Source Wikipedia, article fair use). Cette même source précise que plutôt qu'une liste d'exceptions finies, l'originalité de « l'usage raisonnable » est de fournir aux juges des critères pour déterminer si un usage relève de cette catégorie.

[150]Sur cette affaire, voir Lionel Maurel, « Comment l'affaire Google books se termine en victoire pour le text mining », *S. I. Lex* (blog), 21 octobre 2015 (disponible en ligne sur : http://scinfolex.com/2015/10/21/comment-laffaire-google-books-se-termine-en-victoire-pour-le-text-mining/) (consulté le 20 décembre 2015).

[151]Antoine Oury, « Droit d'auteur : les plans de la Commission européenne pour 2016 révélés », *ActuaLitté*, 9 novembre 2015 (disponible en ligne : https://www.actualitte.com/article/monde-edition/droit-d-auteur-les-plans-de-la-commission-europeenne-pour-2016-reveles/61981) (consulté le 20 décembre 2015).

[152]Voir Christian Handke, Lucie Guibault, Joann-Josep Vallbé, « Is Europe falling behing in data mining ? Copyright's impact on data mining in Academic Research », 2015 (disponible en ligne : http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2608513) (consulté le 3 décembre 2015).

domaine d'un usage raisonnable sans porter préjudice aux ayant-droits, et pour le bien général [153]. Par conséquent, en Europe, tout usage dérogatoire au droit d'auteur, mais ne faisant pas l'objet d'une exception, est de facto illégal. De plus, les deux directives européennes pouvant s'appliquer à cette question, celle sur la défense de la propriété intellectuelle des créateurs de base de données et celle les droits d'auteurs et droits voisins dans la société de l'information, si elles incluaient bien des exceptions au droit d'auteur qui auraient pu permettre la pratique de la fouille de texte, n'imposaient pas leur transposition dans les législations nationales, qui n'a donc pas été systématique[154].

Faute de clarification, une opération de fouille de texte – qui peut donc inclure pour accéder aux informations recherchées une duplication des textes – est toujours illégale en Europe, sauf en cas d'autorisation explicite des ayants-droits. Ce n'est pourtant pas, de la part de l'Union européenne, faute de ne pas avoir perçu l'importance des enjeux. Dans un premier temps, la Commission Européenne avait espéré favoriser le développement du TDM en faisant l'économie d'une réforme du droit d'auteur, par l'encouragement du développement des pratiques contractuelles[155], s'inscrivant dans le prolongement des licences proposées par les éditeurs scientifiques : c'est ainsi que dans le cadre du dialogue Licences pour l'Europe lancé en 2012, un groupe de travail était dédié à l'examen des questions liées au TDM, avec pour objectif explicite de parvenir à un accord sur l'élaboration de licences standards[156]. Le processus avait cependant achoppé rapidement, après que des participants du groupe de travail, dont l'association LIBER, s'en soient désolidarisés pour protester contre un processus exclusivement centré sur la voie contractuelle. Une lettre ouverte a été publiée[157]. Par la suite, conscients de la difficulté, les rédacteurs du rapport de 2014 ont appelé de leurs vœux, au minimum l'addition dans les directives existantes d'une exception au droit de la propriété intellectuelle concernant le text mining, au mieux la rédaction d'une nouvelle directive qui permette la reproduction d'œuvres dans le cadre de travaux à buts non-commerciaux.

Depuis ce rapport de 2014, l'évolution de la situation juridique du TDM en Europe est suspendue au projet de réforme européenne du droit d'auteur, dont le rapport Réda a posé les bases. En novembre dernier, un avant-projet du futur texte a été dévoilé, où l'on peut lire que la Commission européenne a bien l'intention, en 2016, d'introduire une exception au droit d'auteur pour permettre le TDM [158].

[153] Voir Ian Hargreaves (dir.), *Standardisation in the area of innovation and technological development, notably in the field of Text and data mining*, European Commission, Bruxelles, 2014, p. 3 : « In the United States, the 'fair use' defence against copyright infringement appears to offer greater reassurance to researchers than the comparable copyright framework in Europe, which relies upon a closed set of statutory exceptions. » (disponible en ligne : http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf) (consulté le 28 novembre 2015).

[154] Ian Hargreaves (dir.), *Ibidem*, p. 50.

[155] Guillaume Champeau, « Des « licences pour l'Europe » pour éviter une révision du droit d'auteur », *Numerama*, 13 novembre 2013 (disponible en ligne : <http://www.numerama.com/magazine/27501-des-licences-pour-l-europe-pour-eviter-une-revision-du-droit-d-auteur.html>) (consulté le 12 décembre 2015).

[156] Voir, sur le site du dialogue « Licences for Europe », la définition des missions du groupe de travail : « The Group should explore solutions such as standard licensing models as well as technology platforms to facilitate TDM access. » (disponible en ligne : https://ec.europa.eu/licences-for-europe-dialogue/en/tags/text-mining) (consulté le 22 décembre 2015)

[157] Cette lettre ouverte est disponible en ligne : <http://libereurope.eu/wp-content/uploads/Letter_of_withdrawalL4E_TDM_May%2024_4.pdf>)

[158] Antoine Oury, *Ibidem*, 9 novembre 2015.

Conclusion : la situation actuelle en France

En attendant la mise en œuvre de cette réforme européenne du droit d'auteur, et ses transpositions dans les législations nationales, la fouille de texte reste non autorisée, sauf accord contractuel avec les ayants-droits. La permission d'effectuer des opérations de fouille de texte, quand elle est accordée, figure dans les licences d'abonnement conclues avec les éditeurs, et l'opération doit se conduire en respectant les conditions imposées par l'éditeur.

Cette prévalence du point de vue des ayant-droits a, par ailleurs, conduit à la mise en circulation de contrats de licence détaillant des conditions de TDM différentes selon les corpus négociés et les interlocuteurs. Ainsi, dans la licence nationale passée dans le cadre d'ISTEX – pour les archives d'un corpus de 2200 revues scientifiques des origines jusqu'à 2001-, l'éditeur Elsevier a imposé les conditions suivantes pour la fouille de texte sur ses collections : usage obligatoire de l'API de l'éditeur, citation des extraits obtenus par fouille de texte dans la limite de 350 mots, obligation de diffuser le résultat de l'extraction sous une licence non-commerciale[159]. En revanche, les termes de la licence nationale que le même éditeur négociait un an après avec COUPERIN pour les abonnements courants à la Freedom collection, indiquaient pour le TDM les conditions suivantes : limitation du droit de citer les données extraites par TDM à 200 signes, utilisation facultative de l'API d'Elsevier[160]. La régulation contractuelle de l'exercice du TDM a donc pour conséquences de multiplier des régimes spécifiques d'exploitation, avec des conséquences potentiellement absurdes dans la cas de collaborations entre chercheurs ayant accès aux corpus selon des conditions différentes[161].

En l'état, la fouille de texte est donc permise en France pour des contenus sous droits, mais lorsqu'elle est encadrée par des accords contractuels. Bien qu'au fil des négociations avec Elsevier, les conditions imposées par la fouille de texte soient devenues moins drastiques[162], l'absence d'exception juridique ne donne pas de garanties pour l'avenir. Les conditions imposées par les éditeurs au commencement des négociations ont fait craindre la création d'un droit patrimonial sur les données elles-mêmes – puisque Elsevier imposait la mise sous licence des résultats du TDM, alors que les informations visées par la fouille, en tant qu'information, n'appartiennent à personne -, l'imposition de limites arbitraires au droit de diffuser l'information – ainsi la limite imposée à l'origine pour les contenus extraits par content mining, qui allait à l'encontre du droit de courte citation. C'est pourquoi, diverses organisations comme le collectif Savoir Comm1, et les associations professionnelles COUPERIN et ADBU se sont prononcées pour la mise en place d'une exception au droit de la propriété intellectuelle pour permettre la fouille de données.

Toutefois, cette évolution législative n'est pas suggérée sans la mise en place de dispositifs pour donner des garanties aux éditeurs sur la sécurité des opérations de TDM. Ainsi, puisque l'un des principaux problèmes aux yeux des

[159]Le contrat est disponible en ligne, on lit ces restrictions page 12 : <http://www.licencesnationales.fr/wp-content/uploads/Licence-2013-20-ISTEX-ELSEVIER.pdf> (consulté le 20 décembre 2015). Il faut noter toutefois que ces restrictions s'appliquaient à la fouille de texte avant la mise en place de la plate-forme ISTEX, où les données acquises auprès des éditeurs seront dorénavant hébergées.

[160]Les détails de l'accord COUPERIN avec Elsevier sur la Freedom collection n'ayant pas été rendus publics, ces renseignements nous proviennent de l'article de Pierre-Carl Langlais http://scoms.hypotheses.org/276

[161]Pierre-Carl Langlais, « Text-mining : vers un nouvel accord avec Elsevier », *Sciences communes* (blog), 29 octobre 2014, (disponible en ligne sur : <http://scoms.hypotheses.org/276>) (consulté le 20 décembre 2015).

[162]Selon les informations données sur les accords Couperin par Pierre-Carl Langlais (*Ibidem*), la limitation de la citation à 200 signes a, par exemple, été retirée par Elsevier pour les chercheurs bénéficiant de la licence nationale.

éditeurs reste le fait que les opérations de TDM nécessitent la copie intégrale du corpus fouillé, l'idée de tiers de confiances, qui serviraient d'intermédiaire entre ayant-droits et praticiens du TDM, a pu été avancée comme un élément de solution, notamment par l'ADBU[163]. Dans cette configuration, de grands acteurs publics conserveraient des copies des corpus, les mettraient à disposition des chercheurs dans un cadre sécurisé, en limitant ainsi les risques de fuite et de vol des données. On a vu qu'une solution présentant des analogies avait été adoptée, dans un contexte différent, avec le dispositif Data Capsule du HTRC[164], qui lui aussi fournit aux chercheurs un environnement clos pour effectuer des opérations de TDM sur des corpus sous droit, tout en rendant impossible la fuite de ses contenus. Les bibliothèques ou services documentaires sont avancées par l'ADBU comme possibles tiers de confiance : la BNF, en tant que chargée du dépôt légal du Web, pourrait mettre à disposition ce corpus aux chercheurs, et le projet ISTEX pourrait faire de même pour ses propres ressources[165].

[163] Voir Association des directeurs et personnel de direction des bibliothèques universitaires et de la documentation, *Contribution de l'ADBU à la consultation nationale « Ambition sur le numérique » : TDM et Open Access* (non daté et non paginé) (disponible en ligne : <http://adbu.fr/contribution-de-ladbu-a-la-consultation-nationale-ambition-sur-le-numerique-tdm-et-open-access/>) (consulté le 16 décembre 2015).

[164] Voir précédemment, page 43.

[165] Cette deux institutions sont évoquées dans Association des directeurs et personnel de direction des bibliothèques universitaires et de la documentation, *Contribution de l'ADBU à la consultation nationale « Ambition sur le numérique » : TDM et Open Access* (non daté et non paginé) (disponible en ligne : <http://adbu.fr/contribution-de-ladbu-a-la-consultation-nationale-ambition-sur-le-numerique-tdm-et-open-access/>) (consulté le 16 décembre 2015).

CONCLUSION

Quoique les données qu'accueillent les bibliothèques représentent des volumes informatiques sans commune mesure avec ceux générés par l'industrie ou la recherche scientifique de pointe, les collections de documents numériques partagent avec les mégadonnées nombre de caractéristiques : volumes suffisamment importants pour décourager la lecture humaine, hétérogénéité des formats de document, faiblesse ou absence de la structuration, et enfin faible densité informationnelle. Les collections générées par le dépôt légal du Web sont, elles, de plain-pied avec les difficultés propres à l'ère des mégadonnées, puisque Internet est par excellence le lieu natif du big data. Nous espérons donc, à l'issue de ce travail, avoir montré la pertinence de la notion appliquée aux collections numériques des bibliothèques. Ces collections nécessitent donc, et de plus en plus, l'application de techniques dérivées du monde des mégadonnées, à savoir la fouille de texte et de données.

Envisagées comme nous l'avons fait à travers les services apportés aux chercheurs, les instruments de recherche innovants, et les outils d'administration des fonds, ces technologies sont prometteuses, à condition d'en comprendre le fonctionnement, et donc, d'en percevoir les limites. Si nous revenons sur l'assistance que propose la fouille de texte pour le traitement des fonds, cet aspect apparaît clairement. L'indexation intégralement automatisée de documents n'est pas encore évoquée dans les travaux autour de la plate-forme ISTEX, et l'extraction terminologique nécessite toujours, en dernier ressort, l'examen d'un spécialiste pour valider finalement les candidats termes que propose l'algorithme. Pour les professionnels de la documentation, ce que promet la fouille de texte est une assistance, rendue nécessaire par les volumes des fonds, mais dont le travail doit être complété par l'expertise humaine. En tant qu'instruments de recherche, les algorithmes de fouille présentent une réelle pertinence pour permettre des formes d'analyse des corpus relevant de la lecture distante, comme une présentation des thématiques présentes dans des corpus volumineux exprimées avec leur vocabulaire – comme le permettent les algorithmes de modélisations de ce sujet -, mais cependant, ces techniques, qui n'ont pas vocation à remplacer la lecture rapprochée des textes, doivent toujours être proposées comme un point d'accès aux collections, non comme une fin en soi, comme le reproche en a été adressé à Google Ngram Viewer. Par ailleurs, et ce sera le dernier point, ces formes d'analyse des corpus ne peuvent avoir de pertinence que sur des corpus dont les données et les métadonnées répondent à des exigences minimales de qualité.

C'est pourquoi, en conclusion de ce travail, nous dirons que deux types d'obstacles semblent entraver le développement de la fouille de texte sur les collections numériques. D'abord, et pour longtemps, la première difficulté concerne la qualité des données textuelles des collections, ce qui implique que le premier soin pour développer ces usages soit un investissement dans la qualité des données, qu'il s'agisse du plein texte ou des métadonnées. Cette difficulté qualitative prend des formes diverses, plus ou moins graves, et peuvent toucher tous les secteurs des collections : lacunes dans la conversion en mode texte des collections anciennes, médiocrité des métadonnées fournies par les éditeurs, sont autant de problèmes qui peuvent entraver des exploitations globales de collections numérisées pourtant riches. Les difficultés soulevées par les collections acquises dans le cadre des licences nationales ISTEX montrent, d'ailleurs, que la mauvaise qualité des données peut toucher aussi bien des collections récentes, dans la mesure où les documents fournis par les éditeurs ne répondent pas aux exigences

minimales de qualité attendues pour les métadonnées et les données textuelles. Ainsi, nous achèverons en concluant que du côté des bibliothèques, le premier obstacle pour l'extension de la pratique du TDM semble la qualité des données et des métadonnées des collections, qu'il s'agisse des problèmes liés à la conversion en mode texte des ouvrages anciens numérisés, ou les problèmes liés à la pauvreté des données fournies par les éditeurs pour les collections récentes. On peut donc dire qu'en plus du développement des services innovants que nous avons pu évoquer au cours de ce travail, à l'exemple du Hathi Trust Research Center, le premier travail du bibliothécaire pour permettre le TDM semble d'abord cette entreprise d'amélioration de ses données, ce qui représente déjà un défi considérable.

Par ailleurs, il semble clair que l'absence de clarification juridique sur le statut du TDM est bien le second obstacle d'importance pour le développement de ces techniques de « lecture à distance ». En effet, bien que les techniques de text et data mining puissent s'appliquer également aux collections libres de droit, on ne peut ignorer que c'est dans le secteur des publications scientifiques, et pour des besoins de synthèse des recherches récentes encore protégées par le droit d'auteur, que le besoin de recourir à la fouille de texte soit le plus pressant. Le cadre juridique actuel, en empêchant, sauf autorisation explicite des ayant-droits, l'extraction de données, introduit donc une limitation préjudiciable. Il est vrai que dans le cadre contractuel, où le droit de fouille est accordé par les ayants-droits selon des conditions qu'ils édictent, la fouille de texte est cependant possible : toutefois cet état de fait, qui fait dépendre le TDM de l'établissement des licences, n'apporte pas pour l'avenir les garanties souhaitables.

BIBLIOGRAPHIE

BIG DATA

BABINET Gilles. *Big data, penser l'homme et le monde autrement.* Paris, le Passeur, 2015.

DELORT Pierre. *Le big data*. Paris : Presse universitaire de France, 2015.

MAYER-SCHÖNBERGER Viktor et CUKIER Kenneth. *Big data : la révolution des données est en marche.* Paris : Robert Laffont, 2014.

CALDERAN Lisette, LAURENT Pascale, *et alii*. *Big data : nouvelles partitions de l'information / Actes du séminaire IST Inria, octobre 2014.* Bruxelles : de Boeck, 2015.

EKBI Hamid, MATTIOLI Michael, Kouper Inna, Arave G., Ghazinejad Ali, BOWMAN Timothy, RATANDEEP SURI Venkata, TSOU Andrew, WEINGART Scott, et SUGIMOTO Cassidy R.**.** Big data, bigger dilemnas : a critical review. *Journal of the Association for Information Science and Technology,* Août 2015, p. 1523-1545

MEGADONNEES ET SCIENCES HUMAINES

AIDEN Erez et MICHEL Jean-Baptiste. *Culturama : qui n'a jamais rêvé d'avoir lu tous les livres?* Paris : Robert Laffont, 2015.

JOCKERS L. Matthew. *Macroanalysis : digital methods and literary history.* Springfield : University of Illinois Presse, 2013.

GRAHAM Shawn, MILLIGAN Ian et WEINGART Scott. *Exploring big historical data : the historian's macroscope.* Londres : Imperial College Press, 2015. [version préparatoire disponible en ligne : <http://www.themacroscope.org/?page_id=584>] [consulté le 5 décembre 2015].

WILLIFORD Christa et HENRY Charles. *One culture : computationaly intensive research in the humanities and social sciences : a report on the experiences of first respondents to the digging into data challenge.* Council on library and information ressources, 2012. [Disponible en ligne : <http://www.clir.org/pubs/reports/pub151/pub151.pdf>] [Consulté le 5 décembre 2015]

MORETTI Franco. *Distant reading*. Londres : Verso, 2013.

TEXT MINING

JOCKERS L. Matthew. *Text analysis with R for students of Literature (quantitative methods in the humanities and the social sciences).* Lincoln : Springer, 2014.

IBEKWE-SANJUAN Fidelia. *Fouille de textes : méthodes, outils et applications.* Paris : Lavoisier, 2007.

LE TELLIER Isabelle. *Introduction à la fouille de texte* [en ligne]. Paris : Université de Paris 3 – Sorbonne Nouvelle [Consulté le 24 décembre 2015]. [Disponible sur le web : <http://www.lattice.cnrs.fr/sites/itellier/poly_fouille_textes/fouille-textes.pdf>]

WEISS Scholom M. *Text mining : predictive methods for analysing unstructured information*. New-York : Springer, 2005.

LE PROJET ISTEX

BERARD Raymond. *Istex, vers des services innovants d'accès à la connaissance* [en ligne]. Montpellier : ABES, 2012 [disponible en ligne : <http://www.abes.fr/Ressources-electroniques2/Acquisitions/Licences-nationales-ISTEX>].

COLCANAP, Grégory. Istex : un gisement documentaire producteur de connaissances. *Bulletin des bibliothèques de France* [en ligne], n° 1, 2013 [consulté le 31 décembre 2015]. Disponible sur le Web : <http://bbf.enssib.fr/consulter/bbf-2013-01-0066-015>.

BIBLIOTHEQUES ET MEGADONNEES

LEONARD Peter. *Mining large datasets for the humanities.* IFLA, 2014 [disponible en ligne : <http://library.ifla.org/930/1/119-leonard-en.pdf>] [Consulté le 3 décembre 2015]

LEASE MORGAN Eric. Use and understand : the inclusion of services against texts in library catalogs and « discovery systems ». *Library Hi Tech,* Vol. 30, p. 35-39.

ASPECTS JURIDIQUES

Association des directeurs et personnels de direction des bibliothèques universitaires et de la documentation, *Contribution de l'ADBU à la consultation nationale « Ambition sur le numérique » : TDM et Open Access.* [Disponible en ligne : http://adbu.fr/contribution-de-ladbu-a-la-consultation-nationale-ambition-sur-le-numerique-tdm-et-open-access/]

COLCANAP Grégory et PERALES Christophe. *CSPLA, Mission relative au data-mining : l'analyse de COUPERIN et de l'ASDBU*. [Disponible en ligne : http://adbu.fr/wp-content/uploads/2014/04/Audition_CSPLA_TDM_2014_04_04_final.pdf]

HEARGRAVE Ian (dir.). *Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining : Report from the Expert Group* [disponible en ligne : <http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf>]

LANGLAIS Pierre-Carl et MAUREL Lionel. *Quel statut légal pour le content-mining?* [Disponible en ligne : http://www.savoirscom1.info/wp-content/uploads/2014/01/Synthe%CC%80se-sur-le-statut-le%CC%81gal-du-content-mining.pdf]

LANGLAIS Pierre-Carl. *Data-mining : quand Elsevier écrit sa propre loi…*, 8 février 2014 [disponible en ligne :<http://scoms.hypotheses.org/98>] [consulté le 23 décembre 2015]

LANGLAIS Pierre-Carl. *Text mining : vers un nouvel accord avec Elsevier*, 29 octobre 2014 [disponible en ligne :<http://scoms.hypotheses.org/276>] [consulté le 23 décembre 2015]

MAUREL Lionel. *Exploration de données : un environnement juridique en évolution*, 5 décembre 2014 [disponible en ligne : <http://scinfolex.com/2014/12/05/exploration-de-donnees-un-environnement-juridique-en-evolution/>] [consulté le 23 décembre 2015].

INDEX

TABLE DES MATIERES