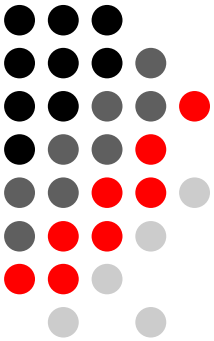
CIRST
Centre interuniversitaire
de recherche sur la science
et la technologie



Note de recherche
2018-01

Numérisation et analyse de documents

Pour une lecture intertextuelle des données massives

Denis Carlier

UQÀM Université de Montréal UNIVERSITÉ LAVAL POLYTECHNIQUE MONTRÉAL UNIVERSITÉ DE SHERBROOKE UNIVERSITÉ TÉLUQ

Pour nous joindre

Téléphone : 514-987-4018
Adresse électronique : cirst@uqam.ca
site : www.cirst.uqam.ca

Adresse postale

CIRST
Université du Québec à Montréal
C.P. 8888, succ. Centre-ville
Montréal (Québec)
H3C 3P8

Adresse civique

CIRST
8e étage
Université du Québec à Montréal
Pavillon Paul-Gérin-Lajoie
1205, rue Saint-Denis
Montréal, Québec

Conception graphique : Cloé Larivière-Jeannotte et Martine Foisy ISBN 978-2-923333-77-9

Notes biographiques

Denis Carlier est membre de l'Institut de recherche en études féministes (IREF) et doctorant en science politique à l'Université du Québec à Montréal (UQAM).

Remerciements

Cette recherche doit beaucoup aux conseils et aux relectures critiques de Jean-Guy Prévost, Benjamin Pillet, Anna Perreault, Alexandre Labonté et Jacques Carlier, ainsi qu'aux commentaires formulés, à propos d'une version préliminaire, par Héloïse Michaud, Catherine Viens, Danielle Coenga Oliveira, James Boyard, Julie-Pier Nadeau, Alioune Ndiaye et Véronique Pronovost, dans le cadre du séminaire doctoral organisé par Jean-Guy Prévost.

Les erreurs et les limites de l'argumentaire demeurent bien entendu les nôtres.

Résumé/Abstract

Cette recherche vise à montrer l'intérêt d'une lecture intertextuelle des ensembles de données massives (*big data*), à même de compenser l'incapacité des approches quantitatives et automatisées à remplir certaines promesses, en particulier face à l'indépassable subjectivité du rapport au texte. On verra ainsi que l'origine commerciale de la notion de données massives participe d'une définition plus marketing que scientifique, à trop vouloir considérer différentes dans leur nature des données qualitatives dont la seule spécificité est en fait leur support numérique. On contestera à ce titre les prétentions à une quelconque représentativité statistique et à une quelconque objectivité dans leur analyse, liées notamment à une confusion dommageable entre la simulation et la réalité simulée. On réfutera de plus la nécessité d'importer en sciences humaines et sociales les définitions des notions de « donnée » ou d'« information » issues du domaine de l'informatique, en invitant à une approche plus conforme au caractère incrémental et itératif du processus d'analyse. On reviendra enfin sur le fétichisme de la technologie sur lequel se fondent les prétentions hégémoniques des humanités numériques, pour leur opposer l'alternative d'une lecture intertextuelle à même d'ouvrir à davantage d'investissement qualitatif des terrains de recherche numériques.

The argument of this paper will be to demonstrate the usefulness of an intertextual reading of big datasets, in order to compensate for the inability from quantitative and automated approaches to fulfill out-of-reach promises, especially regarding the unavoidable issue of subjectivity in the act of confronting one's self to a text. We will show that the business origin of the big data notion is in part responsible for its definition being more of a marketing nature than a scholarly one, in as much as some authors keep looking for a distinctive feature other than its digital medium. We will thus challenge any claim both to statistical representativeness and to objectivity in automated quantitative analysis of big data, partly due to an unfortunate confusion between simulation of reality and the said reality. Moreover, we will reject any necessity to import into the humanities and the social sciences the computing-related definitions of the "data" and "information" notions, in order better to take into account the incremental and iterative essence of the data analysis process. Last but not least, we will offer some thoughts about technology fetishism in the digital humanities, and the hegemonic trend it creates, so as to counter it with our intertextual reading proposal and thus to foster some more qualitative approaches to digital fieldworks.

Table des matières

Introduction

L'avènement des données massives (DM, *big data*), annoncé il y a déjà une décennie par Chris Anderson (2008) dans un éditorial de *Wired*, se trouvait censément annonciateur de la fin des modèles et théories explicatives. Dans un texte devenu le point de départ obligé de toute discussion sur le sujet, le journaliste défend en effet l'idée qu'il serait désormais possible de tout mesurer de manière exhaustive, et affirme en conséquence l'obsolescence des méthodes statistiques. Dans ce nouveau « monde » scientifique prophétisé par Anderson, « les ensembles massifs de données et les mathématiques appliquées remplacent tous les autres outils qui pourraient être mis à contribution », et la méthode hypothético-déductive tout comme la modélisation peuvent être abandonnées sans remords, au profit d'un simple calcul de corrélation (nous traduisons).

L'arrogance de la thèse, qui mène par exemple à considérer avec désinvolture que la physique quantique ne représente qu'un modèle « défectueux », ne serait pas sans faire sourire si elle n'avait pas rencontré un certain succès au cours de la décennie écoulée, y compris en sciences humaines et sociales (SHS). Ainsi, si les prétentions démesurées associées à l'usage des DM ont d'ores et déjà été remises en perspective dans le domaine statistique, notamment après les déboires du projet Google Flu Trend[1] (GFT), un travail similaire demeure à effectuer pour rappeler les apports spécifiques des méthodes qualitatives, un peu trop souvent méprisées au profit d'un positivisme daté, favorisé par l'analyse automatisée sur ordinateur[2].

[1] Ginsberg et al. (2009) justifiaient originellement l'introduction de GFT par l'existence d'une forte corrélation, aux États-Unis, entre l'évolution quotidienne du nombre de requêtes Google pour certains mots-clés et celle du nombre de cas de grippe diagnostiqués. Après une prédiction sous-évaluée en 2009, d'abord attribuée aux spécificités de l'épidémie de grippe A (H1N1) (Cook et al. 2011), un modèle ajusté est par la suite considéré fonctionnel, en dépit d'une légère tendance à la surévaluation, jusqu'en 2013 où le modèle mène à exagérer de 140% le pic de l'épidémie grippale (Butler 2013 ; Lazer et Kennedy 2015). Jusque-là considéré comme une démonstration de l'apport des DM, GFT devient alors à l'inverse un symbole de « l'hubris du *big data* » (Lazer et al. 2014, 1203, nous traduisons) et le projet se trouve abandonné par l'entreprise (Google n.d.). Pour Lazer et al. (2014), par ailleurs optimistes quant aux apports potentiels des DM, la prétention à remplacer les méthodes d'échantillonnage par le recours aux DM n'autorise en rien à ignorer les critères de validité qui s'appliquaient jusque-là.

[2] On laissera pour l'essentiel de côté les *outils d'analyse*, à l'image des logiciels d'aide à l'analyse de données qualitatives (LAADQ, ou CAQDAS pour *computer-assisted qualitative data analysis sofwares*), en concen-

Cette reconnaissance universitaire de la nécessaire diversité des approches méthodologiques semble en effet plus facile face à une critique des humanités numériques (HN, *digital humanities*) formulée sur un même plan quantitatif que face à une critique formulée depuis un point de vue qualitatif, lequel se trouve implicitement relégué à un rang inférieur par une lecture techno-positiviste qui l'assimile à la crainte irrationnelle d'un progrès irrémédiable et salvateur. On montrera pourtant, en insistant sur la question de l'herméneutique, que l'analyse qualitative conserve toute sa pertinence dans le domaine du numérique, et s'avère même par bien des aspects mieux adaptée aux contraintes techniques spécifiques aux DM que ne peut l'être l'approche automatisée en l'état des contraintes techniques. On essaiera d'encourager de la sorte un investissement qualitatif du terrain numérique, en nous concentrant sur une stratégie de recherche par l'étude de documents, soit l'exploitation la plus évidente des données massives envisageable du point de vue des SHS.

On s'intéressera donc tout d'abord à la notion de données massives, pour en contester la spécificité et les requalifier comme des données qualitatives. Puis on verra que l'approche quantitativiste des HN n'est pas aussi nécessaire que ne le laissent à penser ses défenseurs et enfin qu'une « lecture intertextuelle » s'avère particulièrement adaptée pour la recherche documentaire en ligne.

1. Les données massives sont des données qualitatives

D'une manière comparable celle de « gouvernance » (Crowley 2003), la notion de « données massives » se trouve aujourd'hui omniprésente dans la littérature scientifique autant que dans la production journalistique (Kitchin 2014b, 67), malgré le fait que sa définition demeure souvent très floue. Malgré certains efforts pour plus de précision dans l'emploi du terme, la notion comporte en effet fréquemment un aspect messianique assez prononcé. Plutôt que de

trant nos critiques sur les *principes méthodologiques*. L'utilisation de logiciels à des fins d'analyse de discours nous semble ainsi pertinente tant et aussi longtemps qu'elle reste en cohérence avec la méthode d'échantillonnage et la méthodologie générale retenues, et qu'elle ne mène pas à des quantifications hors de propos. On se rallie ainsi à la position normative de Christophe Lejeune (2017, 223-224), résumée par la formule « frugalité informatique et audace intellectuelle », qui suppose que « [t]ravailler sans logiciel devrait […] toujours rester une option ».

données massives, c'est alors plutôt de « *big data* » qu'il est question, c'est-à-dire d'une idée synthétisée par un terme-parapluie inutile à traduire ou à définir, porteuse d'un « progrès » qu'« [i]l est vain de chercher à retenir », d'une « révolution » dont « l'avènement […] est proche », ainsi que le pense par exemple Gilles Babinet (2015, 14), défenseur français du numérique (*digital champion*) pour l'Union européenne (UE).

Mais même à laisser de côté ce catéchisme zélé pour se concentrer sur des auteurs plus sérieux dans leur réflexion sur les DM, on constate que l'origine commerciale mal connue de la notion continue de structurer sa théorisation académique. En effet, on démontrera ici que les DM ne possèdent pas les spécificités qu'on leur prête et l'on reviendra sur la nécessité de ne pas céder au fétichisme de la technologie dans l'approche théorique des données. On montrera ainsi que les DM se trouvent tout à fait accessibles à l'analyse qualitative en sciences humaines et sociales (SHS) et ne nécessitent donc pas forcément d'automatisation de l'analyse comme il est souvent répété sans démonstration.

1.1 Une notion apparue par contagion

L'expression de « *big data* », qu'on choisit de traduire ici par « données massives[3] » renvoie en premier lieu à une notion utilisée dans les cercles académiques et professionnels avant d'être conceptuellement définie (Diebold 2012). Si cette caractéristique n'invalide pas l'affirmation de la spécificité des DM, elle permet du moins de comprendre en partie comment un certain nombre de particularités ont pu leur être prêtées exagérément.

Les « trois V », généralement utilisés pour définir les DM, à savoir le « volume », la « vitesse » et la « variété », sont ainsi tout d'abord théorisés en 2001 par le consultant Doug Laney (2001) pour le cabinet d'étude Meta Group, à propos de ce qu'il nomme alors les « données 3D » (nous traduisons). L'objectif est commercial et n'a qu'un rapport distant avec l'analyse des données massives puisqu'il s'agit d'« aider [les] clients à saisir et, de façon plus

[3] Ce choix a l'avantage de faciliter les traductions depuis l'anglais puisqu'elle permet par exemple de transmettre le parallélisme de construction entre la notion de « données massives » et celle de « données ouvertes » (*open data*). L'expression « *massive data sets* » a par ailleurs pu être utilisée avant que « *big data* » ne s'impose (e.g. Kettering et al. 1997).

importante, à faire face [aux] défis » associés au développement des progiciels de gestion intégrée (PGI, *enterprise resource planning*) du début des années 2000[4] (Laney 2012, nous traduisons). Un des usages premiers de la notion demeure d'ailleurs aujourd'hui encore commercial, ce qui n'aide pas à en clarifier la définition. Edd Wilder-James (2012) note par exemple l'ajout incessant par les « marchands de logiciel » de nouveaux « V » à la liste déjà existante, sans pour autant rien ajouter aux éléments identifiables comme spécifiques aux DM[5] (nous traduisons). Cette catégorisation « trop pleine de langage commercial gestionnaire » est également critiquée par Emma Uprichard (2013), qui estime qu'elle mène à « passe[r] à côté de ce qui est important pour [elle] comme chercheuse de sciences sociales, qui est de rendre plus visible [l'endroit] où se situe le réseau de pouvoir ». Malgré la contradiction entre cette inflation définitionnelle à des fins marketing et le principe du rasoir d'Occam, des tentatives de rapprochement avec l'approche scientifique ont lieu, par exemple aux États-Unis à l'initiative du National Institute of Standards and Technology (NIST), dans un effort de consultation des chercheur·e·s, des fonctionnaires et des professionnel·le·s (Grady [2015] 2017).

S'en suit une relative « contagion » des approches, malgré des tentatives de proposer une définition proprement scientifique. Le signe le plus évident de cette tendance est sans doute l'omniprésence dans la littérature d'une prémisse implicite selon laquelle l'information serait

[4] La note de Laney (2001) est rédigée spécifiquement en rapport avec la gestion des données liées au commerce électronique, en abordant successivement a) le coût de stockage d'un « volume » plus important de données lié au développement de l'activité, b) la nécessité de maximiser la « vitesse » d'accès au site Internet et à ses données à des fins concurrentielles – l'accès dit à « haut débit » commence alors à peine à se développer – et c) les stratégies de mise en compatibilité des données malgré leur « variété », i.e. malgré leur caractère non structuré.

[5] On compte aujourd'hui facilement sept « V », avec « le volume, la vitesse, la variété, la variabilité, la véracité, la visualisation et la valeur » (DeVan 2016, nous traduisons). La « variabilité » se réfère à une valeur possiblement différente selon le moment ou le contexte pour une même donnée, par exemple un changement de la signification associée à un mot. La « véracité » renvoie au risque potentiel associé à l'utilisation de données fausses ou corrompues. La « visualisation » renvoie à la possibilité de représenter graphiquement les données. Enfin, la « valeur » est l'objectif commercial de maximisation du profit économique (McNulty-Holmes 2014 ; DeVan 2016). Aucune de ces nouvelles composantes n'est spécifique aux DM et c'est donc sur un ton moqueur que Wilder-James (2012) propose la « vérisimilitude » comme « 8^{e} V des données massives que tout le monde avait manqué ».

une ressource dotée d'une valeur intrinsèque, dans une vision directement issue d'une approche commerciale de l'informatique[6]. L'expression même de *data mining*, traduite notamment par « forage », « fouille », « prospection » ou « extraction », reprise telle quelle dans la littérature scientifique, renvoie d'ailleurs à un imaginaire industriel et capitaliste. John Atkinson (2009) propose ainsi un parallèle direct entre la recherche d'« or » et la recherche d'« information », chacun·e étant considéré·e comme un minerai obtenu par le biais d'une extraction et de transformation à grande échelle, qui confère aux « pépites » (*nuggets*) une valeur intrinsèque, laquelle participe à l'augmentation d'un capital accumulé de connaissances (nous traduisons).

De manière corollaire, cette conception de l'information comme ressource encourage à quantifier l'inquantifiable, jusqu'à atteindre des sommets d'absurde comme lorsque l'Open Data Center Alliance (ODCA 2012, 6) affirme très sérieusement que les 2,5 pétaoctets de données commerciales produites chaque heure par Walmart représentent « 167 fois l'information contenue dans l'ensemble des livres de la bibliothèque du Congrès[7] » (nous traduisons). Le recours à une avalanche de chiffres sans signification au regard de leur valeur absolue n'est hélas pas l'apanage des productions à des fins industrielles, puisque le géographe irlandais Rob Kitchin (2014a, 2) cède par exemple lui aussi à ce travers, de même que le philosophe italien Luciano Floridi (2010, 22).

Mais l'élément de « contagion » porteur des conséquences les plus importantes pour notre défense d'une approche qualitative des DM s'avère le préjugé selon lequel les « processus

[6] Par exemple, pour la société de progiciels Oracle (2013), l'apport de l'analyse automatisée des DM est l'élargissement de l'utilisation des données pour des « décisions commerciales » et donc la création de valeur, en ne limitant plus l'analyse aux données « entreposées dans des bases de données relationnelles » comme les « données de transaction », mais en intégrant également des données moins structurées, qu'il s'agisse de blogs, de courriels, de photographies, etc.

[7] L'OCDA compare ici des choux et des carottes. D'une part, l'ensemble du catalogue de la bibliothèque du Congrès n'a pas été numérisé. D'autre part, quand même ce serait le cas, il n'existe aucune corrélation entre la taille d'un fichier et le sens associable à son contenu. Une différence de format ou de résolution suffira ainsi à faire varier la taille d'un fichier de manière significative, sans impliquer une différence majeure dans l'information transmise, au moins dans le cas des documents écrits. Enfin, les ouvrages peuvent être interprétés individuellement là où les données de transaction, utilisées dans le seul objectif de maximiser les profits, n'ont d'intérêt à cette fin qu'une fois agrégées pour en tirer des patterns.

ou outils traditionnels » d'analyse de l'information ne « peuvent pas » être appliqués aux DM (Zikopoulos et al. 2012, 3, nous traduisons). La *difficulté* à prendre en compte de manière automatisée les données non structurées, qui justifie à l'origine la distinction entre les DM et le reste des données, se trouve en effet désormais présentée à l'inverse comme un argument *en faveur* de l'analyse automatisée en donc d'un changement épistémologique profond en SHS. C'est par exemple le cas lorsque Jeffrey Schnapp et al. (2009, 13) rejettent toute possibilité d'un « virage numérique qui pourrait d'une quelconque manière laisser intactes les sciences humaines », c'est-à-dire ne pas impliquer d'analyse automatisée des DM, nouvelle matière première des SHS (nous traduisons). Or, la « capacité d'englober des données hétérogènes » constitue justement une des forces de l'approche qualitative (Pires [1997] 2007, 74), et les données semi- ou non-structurées[8] représentent même le plus souvent la matière première de l'analyse qualitative, en ce qu'elles constituent des « données de première main », moins sujettes aux biais de ses interprètes successifs (Pires [1997] 2007, 13).

Pour Alvaro Pires ([1997] 2007, 73), le recours à ce « matériau empirique qualitatif, c'est-à-dire non traité sous la forme de chiffres » est même l'élément le plus directement distinctif d'une approche qualitative, et l'on peut remarquer que la notion de données massives se trouve en premier lieu définie en fonction d'une approche quantitativiste, qui ne remet pas en cause ce caractère proprement qualitatif des données mais l'interprète au regard de ses propres limites. Les théoriciens des DM ne font en effet que se donner pour défi de structurer les données qualitatives sous une forme quantifiée (cf. 2.1).

[8] L'expression de « données non-structurées » concerne en informatique ce que Ted Nelson (2014) appelle des « blocs de données » (*lump files*, nous traduisons), c'est-à-dire des ensembles indifférenciés de données stockés dans des fichiers uniques. Le contenu lui-même, difficilement accessible à l'analyse automatisée, peut donc à l'inverse être considéré très structuré au niveau sémantique et/ou syntaxique. Les données « semi-structurées », en comparaison, contiennent déjà des métadonnées, ou bien ces dernières peuvent en être déduites automatiquement, comme dans le cas des courriels (Feldman et Sanger 2007, 3-4). Enfin, les métadonnées sont des « données structurées à propos d'un objet[,] qui supporte les fonctions associées à l'objet en question » (Greenberg 2003, 1876, nous traduisons) : à un livre sera par exemple associé un titre, un·e ou plusieurs auteur·e·s, une date de publication, etc.

1.2 L'absence de spécificité des données massives

La non-spécificité des DM s'avère assez aisément démontrable, mais cette hypothèse se trouve simplement ignorée dans les définitions proposées dans la littérature scientifique, dont les auteur·e·s prennent pour acquis ce qu'il conviendrait en fait de démontrer. On reprendra ici, pour montrer l'existence de ce biais de confirmation, la réflexion de Rob Kitchin, l'une des plus précises et des plus détaillées formulées en SHS, mais qui souffre de n'avoir pas fait le bilan critique de l'origine de la notion et semble surtout guidée par une volonté de transposer aux données elles-mêmes une distinction de nature normative et méthodologique.

Kitchin (2014a, 2; 2014b, 68) reprend en effet comme base à sa définition les « trois V » de Laney (2001), qui se trouvent toutefois considérés désormais à des fins analytiques plutôt que gestionnaires, non sans conséquences. La notion de « volume » en perd tout d'abord sa pertinence, puisque l'analyse des données n'a pas à se préoccuper des coûts associés à la constitution et à la gestion d'une base de données. En d'autres mots, la persistance de la croyance en une valeur intrinsèque des données empêche Kitchin de remplacer la notion de volume par celle d'exhaustivité, qu'il introduit comme complémentaire. Dans un deuxième temps, la notion de « variété » évolue légèrement, puisque Laney était préoccupé par des enjeux de convertibilité entre données et insistait sur la diversité interne aux données non-structurées, tandis que Kitchin insiste plutôt sur la coexistence de données structurées et non-structurées pour appuyer un élargissement des capacités d'analyse – oubliant au passage l'existence de l'approche qualitative[9]. Enfin, la « vitesse » change complètement de signification, puisqu'au lieu de la vitesse d'accès aux données, le terme renvoie désormais chez Kitchin (2014b, 76) au caractère « beaucoup plus continu » qu'auparavant de la « génération des données ».

Kitchin (2014b, 68) ajoute également plusieurs caractéristiques, issues d'une revue de la littérature existante. Certaines paraissent pourtant d'emblée inadéquates, soit qu'elles soient

[9] Reprenant à ce propos l'affirmation de l'OCDA (2012, 7) selon laquelle, jusqu'à l'émergence des données massives, « les données non structurées étaient soit ignorées, soit au mieux utilisées inefficacement », Kitchin (2014b, 77) participe activement à l'invisibilisation de l'approche qualitative (cf. 3.3).

non spécifiques aux DM comme la nature « relationnelle[10] » ou « extensible[11] » des ensembles de données, soit qu'elles conservent une préoccupation technique à rebours du reste du propos comme pour l'enjeu de la « scalabilité » associée à la possibilité d'une extension rapide du nombre de données. Semblent enfin mériter plus d'attention le critère d'« exhaustivité », soit l'ambition de faire coïncider un échantillon *n* avec une population *N*, et les critères de « résolution » et d'« indexation », soit l'accessibilité à un niveau élevé de détail et à une identification spécifique à chaque individu, objet ou document.

Or, ces critères potentiellement pertinents de vitesse, d'exhaustivité, de résolution et d'indexation concernent certes des « caractéristiques ontologiques » des données (Kitchin et McArdle 2016, nous traduisons) mais ils ne s'appliquent pas pour autant à tous les cas généralement considérés comme relevant des DM, chaque critère comportant des contre-exemples. Ils enferment dès lors les DM dans une définition ajustable aux circonstances, plus marketing ou rhétorique que scientifique.

Pour commencer, le caractère continu de l'agrégation des données (« vitesse ») n'est pas constatable pour l'ensemble des données généralement considérées comme massives. Si c'est par exemple le cas pour des contenus spécifiquement numériques – publications ou interactions (« likes », « retweets », etc.) des réseaux sociaux, requêtes sur un moteur de recherche, etc. – ou pour des contenus produits pour une publication à la fois numérique et non-numérique – articles de presse, articles scientifiques, etc. –, ce caractère continu est beaucoup moins évident dès lors qu'il s'agit d'importer des contenus d'une origine « étrangère » au médium numérique. Google Street View (GSV), pensé comme un projet de « numérisation » du réseau routier et du cadre bâti environnant par un biais photographique, met ainsi ses

[10] On peinera en effet à trouver une quelconque activité intellectuelle qui ne repose pas sur la mise en relation d'ensembles de données. L'exemple des campagnes présidentielles de Barack Obama en 2008 et en 2012, cité par Kitchin (2014b, 74), laisse à penser que c'est l'ampleur et/ou la systématicité de l'effort de mise en relation de données hétérogènes qui est en jeu, mais c'est là un enjeu de méthode et non de nature des données.
[11] L'« extensionnalité », soit la possibilité d'ajouter de nouveaux champs, n'est en rien spécifique aux données elles-mêmes. La rigidité des méthodes statistiques, visée par Kitchin, se retrouve ainsi contournable en partie, par exemple par le biais d'un indice longitudinal de préférences publiques (*public policy mood*) qui agrège des séries de questionnaires et infère des données manquantes pour établir une tendance (Stimson, Tiberj et Thiébaut 2010 ; Tiberj 2014).

données à jour tous les ans environ pour les grandes métropoles occidentales[12], ce qui demeure très loin du temps réel ou quasi temps réel de Kitchin, malgré la tendance continue à l'extension de la couverture géographique du service[13].

Le double-critère de « résolution » et d'« indexation » unique souffre également d'exceptions et nous semble surtout répondre à la recherche d'éléments distinctifs d'avec les données de recensement. En d'autres mots, les données massives semblent davantage définies en fonction de divergences normatives quant aux méthodes prônées que de différences dans la nature des données visées par l'analyse. Le reproche formulé par Kitchin (2014a, 2) selon lequel les données du recensement seraient « généralement assez grossières (*coarse*) dans leur résolution » en ce qu'elles demeurent limitées au niveau d'« aires locales » ou de « comtés » sans inclure les « individus et foyers » nous parait en effet contestable pour trois raisons (nous traduisons). La première est que la « granularité » des données, c'est-à-dire leur niveau de détail, se trouve limitée pour des raisons éthiques et politiques et non pour des raisons techniques, puisque les données de ce que Statistique Canada ([2016] 2017) nomme l'aire de diffusion (AD) sont constituées de l'agrégation des données de recensement pour chaque foyer. La deuxième raison en est que les DM ne sont pas non plus exemptes de reproches du point de vue de ce même critère : l'indexation unique des données de Google Books (GB) sera par exemple impossible, puisque les données brutes téléchargeables de l'outil Google Ngram Viewer (GNV) n'incluent pas les métadonnées des livres analysés, pour des raisons de copyright qui limitent de fait la portée des analyses envisageables[14] (Google [c2009] c2012 ; Koplenig 2015). Enfin, la troisième raison est que la définition de la « résolution »

[12] À prendre pour référence l'hôtel de ville de Montréal, les données de Google Street View ont ainsi été mises à jour onze fois entre septembre 2007 et décembre 2017. De même, on peut relever huit mises à jour sur la même période au niveau de l'Empire State Building à New York et huit mises à jour entre juin 2012 et décembre 2017 au niveau de la tour Eiffel à Paris.

[13] Cette extension continue connait quelques exceptions notables comme l'Allemagne ou l'Inde, respectivement pour des arguments de respect de la vie privée et de sécurité (Murphy 2011 ; Ohja 2017).

[14] Ces données sont par ailleurs incomplètes et n'ont pas été mises à jour depuis 2012. Les cinq millions de livres du corpus initial de GNV, soit un tiers du corpus de GB à ce moment-là ont été sélectionnés en fonction de « la qualité de leur OCR et de leur métadonnées » et subdivisés selon la langue détectée, en excluant également les périodiques (Michel et al. 2011, 176, nous traduisons) – l'« OCR » étant la reconnaissance optique des caractère (*optical character recognition*).

des données pourra varier fortement selon l'objectif pour lequel on voudra mobiliser ces dernières. GSV permettra par exemple de répertorier de manière virtuellement exhaustive les bâtiments contigus à une rue, mais présentera des limites pour identifier les numéros de rue affichés sur ces mêmes bâtiments. À défaut d'une définition absolue, les données de GSV seront donc, comme le chat de Schrödinger, à la fois massives et non-massives tant qu'on ne définit pas de protocole d'« observation[15] ».

Enfin, la mise à jour des données en quasi temps réel n'implique pas forcément d'approcher l'« exhaustivité », en particulier en raison des coûts associés à cet objectif. On peut ici penser en particulier au cas des archives de presse, que résume bien celui de *La Croix*. Le quotidien catholique français a ainsi numérisé ses archives jusqu'en 1996, et n'ambitionne pas en l'état de le faire au-delà de cette date, n'augmentant de fait l'amplitude temporelle de ses archives que par la seule action du temps (quinze ans d'archives en 2011, vingt ans en 2016, etc.). Et ce n'est que par un financement public via la Bibliothèque nationale de France (BNF) et sa plate-forme numérique Gallica que la numérisation des archives a pu commencer en 2005 pour les numéros publiés entre 1883 et 1944 (Giuliani 2005). Si l'on peut toujours prétendre que le rapport « n/N » augmente en conséquence chaque jour et rapproche de l'exhaustivité le volume numérisé des archives, cette affirmation ne comblera pas pour autant le vide d'un demi-siècle dans l'accessibilité numérique aux numéros du journal, lequel ne s'est pas résorbé depuis l'annonce du partenariat il y a douze ans. Et malgré quelques contre-exemples comme les archives du *Monde* ou du *Nouvel Observateur* (aujourd'hui *L'Obs*) numérisées jusqu'au premier numéro, soit respectivement depuis 1945 et depuis 1964, les articles de presse de la période comprise entre la fin de la Seconde Guerre mondiale et le milieu des années 1990 demeurent très largement inaccessibles en ligne pour la France, même de façon

[15] Google a d'ailleurs pris en compte la question avec une évolution du dispositif de prise de vue en 2017, « première amélioration majeure en huit ans » (Simonite 2017b, nous traduisons). L'objectif y est en particulier d'améliorer la résolution des photographies pour faciliter l'analyse automatique du contenu des images photographiées et rendre mieux déchiffrable aux algorithmes des indications comme les noms de commerce et leurs heures d'ouverture affichées. Ces données à la granularité plus fine sont en retour censées permettre d'améliorer les possibilités de recourir à l'apprentissage machinique (*machine learning*) pour développer un décryptage automatique, par comparaison de données, d'abréviations sur les panneaux de signalisation. Autant d'évolutions qui laissent à penser que la version actuelle de GSV pourrait bien être rétrospectivement considérée comme du « *small data* ».

payante. Le critère de quasi exhaustivité est donc de fait dénué de sens dès lors qu'on analyse des données historiques.

1.3 Une absence de représentativité statistique

Ce constat d'une limite au critère de quasi exhaustivité est extensible à la question de la représentativité statistique des données numériques disponibles, à rebours du postulat selon lequel le volume des DM suffirait à créer leur représentativité. Il est tout d'abord à relever que, même à accéder à l'ensemble des publications d'un réseau social, il demeure illusoire d'espérer produire une analyse représentative d'autre chose que de l'utilisation qui est faite des possibilités d'expression offertes par cette plate-forme[16], au point que Laëtitia Émerit-Bibié (2016) en vient à revendiquer un « droit à la non-représentativité » de son analyse des contenus Facebook, rejoignant de fait une démarche plus qualitative. Cette représentativité limitée est de plus illusoire, puisque les données ne sont en fait le plus souvent pas complètement accessibles : outre une option gratuite au rabais, Twitter tarifie ainsi entre 150 $ et 2 500 $/mois différents niveaux d'accès « premium » à ses données des trente derniers jours, un prix bien inférieur à l'accès supposément complet des forfaits pour entreprise (Perez 2017 ; Dignan 2017 ; Twitter 2017). Mais même à pouvoir financer un accès complet, seront tout de même d'emblée exclus les tweets provenant de comptes protégés, soit environ 8% des comptes selon une estimation de 2010, et seront à l'inverse considérés sur le même plan des comptes personnels, collectifs, d'entreprise ou de bots. De plus, de nombreux problèmes techniques dans la base de données centrale mènent à des suppressions régulières de tweets et le peu d'informations publiques sur le mode d'échantillonnage limite la portée de

[16] Les études d'opinion fondées sur des données de Twitter abandonnent d'ailleurs la prétention à une représentativité statistique pour lui substituer « l'hypothèse selon laquelle les individus présents sur Twitter, plus instruits que la moyenne, p[euvent] être considérés comme des "leaders d'opinion" qui influence[nt] indirectement l'ensemble de la population » (Boyadjian 2014), soit la théorie de la communication à double-étage de Paul Lazarsfeld (cf. Katz 1957). Dubitatif, Julien Boyadjian (2014) constate en particulier une surreprésentation des jeunes, des hommes, des étudiants et des professions intellectuelles supérieures, puis note que cette non-représentativité est encore renforcée dans l'expression publique d'opinions politiques par l'impact du genre, de l'âge, du niveau de diplôme et de la profession, en particulier en dehors des périodes de forte politisation comme lors des campagnes électorales majeures. Enfin, le politiste français relève une surreprésentation des partis de gauche parmi les intentions de vote des membres de son panel. Autant de raisons pour lui de relativiser la capacité à « expliquer de manière sociologique la corrélation supposée entre tweets et votes ».

l'interprétation. En un mot, il est important de distinguer « données massives » et « données totales » (*whole data*) (boyd et Crawford [2011] 2012).

En plus de la représentativité par rapport à la population générale, se pose également, pour les objets d'origine extra-numérique, la question de la représentativité par rapport à la population disponible à la numérisation. Comme on y a déjà fait allusion avec la presse française, le catalogue de documents disponible en ligne dépend en effet des politiques de numérisation ainsi que de financements parfois fluctuants[17]. Peuvent pêle-mêle être pris en compte l'état de conservation des documents, leur caractère patrimonial, les sources privées de financement, l'état de conservation, la complexité technique de la numérisation d'un format donné, etc., et Pierre-Carl Langlais (2017) constate par exemple, dans le cas de Gallica et pour les livres publiés avant 1900, une grande disparité dans les choix de numérisation selon l'année de publication, avec 31% du catalogue numérisé pour l'année 1731 mais 3% seulement pour l'année 1530, pour une moyenne de 18% du catalogue numérisé et une tendance générale à privilégier le XVIII^{e} siècle[18].

Ce constat remet en cause la prétention à abandonner les standards de rigueur des approches statistiques, mais il tend à être pris un peu à la légère par un certain nombre d'auteurs. Viktor Mayer-Schönberger et Kenneth Cukier (2013, 29) confondent par exemple volontairement « rassembler autant [de données] que possible » (volume) et les rassembler « tout[es] » (ex-

[17] La numérisation des titres de la presse quotidienne régionale (PQR) constitue par exemple « un axe prioritaire de la politique de conservation, de diffusion et de valorisation du patrimoine régional ancien » de la région Midi-Pyrénées (aujourd'hui intégrée à la nouvelle région Occitanie), notamment du fait de « la mauvaise qualité des matériaux entrant dans leur fabrication ». C'est donc vers la plate-forme Rosalis de la bibliothèque de Toulouse qu'il faudra se tourner pour consulter les exemplaires numérisés de *La Dépêche du Midi* ou du *Midi socialiste* (Courtial et Lavigne 2011). À l'inverse, les coupes budgétaires à Bibliothèque et archives nationales du Québec (BANQ) ont mené en 2017 à un ralentissement du programme de numérisation, au risque de pertes complètes pour le matériel audiovisuel conservé sur bandes magnétiques (Lalonde 2017).

[18] Les données brutes disponibles au téléchargement sont celles de la base de données de Data BNF, qui ne correspond pas exactement au contenu du catalogue de la BNF, mais il s'agit de fait de la référence choisie par l'institution publique, les deux ensembles ayant vocation à converger (BNF c2013 ; Simon 2015). Langlais teste aussi une seconde métadonnée, à savoir la pagination, plus élevée que la moyenne annuelle pour les livres numérisés datant d'avant 1900, et suggère que ces différences se retrouveront pour les autres métadonnées accessibles.

haustivité), citant pour justifier cet écart à une prémisse scientifique de base (le tout est différent de la partie) l'exemple Google Flu Trends… dont l'échec s'explique justement par ce raccourci (nous traduisons). De même, Kitchin (2014b, 72) définit l'exhaustivité par la propension à approcher *N* ou à utiliser « au moins de bien plus larges échantillons que ceux traditionnellement employés dans des études sur des données restreintes (*small data*) ».

2. Données massives et simulation : « La cuillère n'existe pas » (Wachowski et Wachowski 1999)

On a donc vu dans la première partie que l'origine commerciale de la notion de données massives confère à celle-ci une dimension plus marketing que scientifique, notamment par l'insistance à chercher une spécificité aux DM. De plus, cette notion arrivée par contagion dans le domaine scientifique perd de fait toute spécificité dès lors que l'on met au test les critères définitionnels proposés. Enfin, toute prétention à la représentativité statistique peut également se trouver contestée, y compris pour des données d'origine spécifiquement numérique. Or, comme on le verra à présent, le postulat erroné conférant une spécificité exagérée aux DM n'est pas sans répercussions sur les méthodes de recherches en environnement numérique, puisqu'il facilite un positivisme analytique qui ignore toutes les exigences d'autodistanciation critique associées à l'interprétation des données (Bourdieu 1978 ; Harding 1992).

2.1 Les documents numériques et la confusion entre contenu et support

Remettre en cause la validité des critères de distinction mobilisés par Kitchin mène logiquement à contester la spécificité des données massives, et à les considérer en premier lieu comme des *données qualitatives*. Tout comme une tente individuelle renommée « Pause Pod » pour plaire à un public de cadres demeurera en premier lieu une tente individuelle (Schwedel 2017), la révolution des données massives est donc fondamentalement un « réem-

ballage » informatique des données qualitatives. Et l'analyse des données massives nécessitera de fait une quantification préalable des données qualitatives, c'est-à-dire leur mise en conformité avec un modèle mathématique abstrait, leur « datafication[19] ».

Si les DM se trouvent porteuses d'une caractéristique première en comparaison avec les données statistiques, c'est en effet d'être de nature qualitative, et en conséquence de n'avoir pas été rassemblées à des fins de quantification et de devoir subir une transformation préalable à toute analyse automatisée. Le caractère « massif » n'est alors pas attribuable aux données elles-mêmes mais à leur mise en lien informatisée, ce qui n'implique donc en rien que leur contenu soit spécifique au médium numérique. La numérisation des articles scientifiques, par exemple, ne crée pas en elle-même de connaissance nouvelle, mais en facilite la production *du fait* de l'accessibilité accrue créée par la mise en relation de ces contenus dans des bases de données océrisées, c'est-à-dire dans des bases rassemblant des données informatiques dont le contenu textuel a été transcris par reconnaissance optique des caractères (OCR) et se trouve potentiellement exploitable de façon automatisée. L'ambition associée à l'analyse des DM en SHS renvoie alors à une volonté de pousser l'automatisation un cran plus loin, de l'océrisation et l'indexation vers l'analyse de contenu.

La principale distinction entre les données massives et le reste des données qualitatives est donc leur caractère numérisé, qui ne doit pas être confondu avec leur contenu[20]. Certes, tout

[19] Loin de mener à une fin des modèles comme le croit Anderson (2008), l'analyse automatisée des DM est donc guidée par un projet de modélisation mathématique de l'environnement social, géographique, politique ou autre. Or, pour citer Benoît Mandelbrot ([1975] 1995, 7), dans certains « domaines la réalité se révèle être si irrégulière[ ]que le modèle continu parfaitement homogène » que cette datafication vise à émuler « déçoit, et qu'il ne peut même pas servir comme première approximation » – le mathématicien français fait ici référence aux objets fractals, qui exigent de considérer la possibilité d'un nombre non entier de dimensions, à rebours du modèle géométrique euclidien. Même à accepter l'idée d'une continuité directe entre sciences naturelles et formelles (SNF) d'une part et sciences humaines et sociales de l'autre, une prudence s'impose donc quant aux angles morts de toute modélisation, appelant à la fois à la reconnaissance d'une nécessaire diversité méthodologique et à l'abandon de toute prétention à englober par une modélisation unique l'ensemble de la réalité observable. En l'occurrence, on insistera en troisième partie sur l'un des apports les plus pertinents de l'approche qualitative, celui de l'interprétation subjective des contenus sémantiques.

[20] On n'abordera pas directement ici l'enjeu des différences d'interprétation d'un même contenu selon le parcours intellectuel individuel, selon le contexte social, culturel ou historique, selon la position personnelle dans la structure des rapports de domination, etc., qui interdit toute interprétation unique sauf à posséder un capital social et culturel suffisant pour imposer (momentanément) une interprétation canonique (Bourdieu 1979 ; Hall [1980] 1997 ; Haraway [1988] 2007 ; Harding 1992 ; Pollock [1999] 2007).

support « participe de l'expérience d'appréhension du texte par le lecteur » (Thérenty 2009) et « il n'est pas de texte hors le support qui le donne à lire » (Chartier 1988, 16), mais il convient de ne pas céder au *fétichisme de la technologie*, c'est-à-dire à l'attribution à tout contenu numérisé d'une spécificité qu'il ne possédait pas avant d'être numérisé. La question du support s'avère tout d'abord indépendante de celle du numérique, puisqu'une lecture sera par exemple différente selon qu'elle se fait « à haute voix, pour soi ou pour les autres, […] en silence, […] [en son] for privé », etc. (Chartier 1988, 20). Comme l'analyse Marshall McLuhan ([1964] 2013, 12-13), « le "contenu" de chaque médium est toujours un autre médium[21] » et à chaque médium correspond un contenu spécifique, qui « ne pourrait exister sans » celui-ci (nous traduisons). En l'occurrence, le langage est un contenu de l'écriture, elle-même contenu du texte imprimé (*print*), lui-même contenu du livre, lui-même contenu du fichier numérique, lui-même contenu de la base de données ou du réseau qui y donne accès. Autrement dit, le contenu spécifique de la base de données, celui qui « ne pourrait exister sans » son médium, est donc le *fichier numérique*, la base de données ou le réseau mettant en relation ces contenus jusque-là indépendants[22]. La différence est d'autant plus importante que le support numérique demeure aujourd'hui le plus souvent centré sur l'imitation du support imprimé[23], tout comme le cinéma des débuts a, pendant une certaine période, imité la forme théâtrale avant de s'en émanciper, notamment par les outils du cadrage et du montage[24] (Amiard-Chevrel 1990).

[21] McLuhan ([1964] 2013, 12) évite la régression infinie en plaçant au bout de la chaîne de contenus le processus de pensée non-verbal, auquel il ne s'intéresse pas directement.

[22] « Un fragment de poterie reste un fragment de poterie dans la collection de l'archéologue, mais il sera commenté, en fonction de son insertion dans un ensemble d'autres objets », estime Dominique Cotte (2004, 35) pour distinguer « objet » et « document ». Si l'on tendrait davantage à suivre ici l'approche de Barthes ([1973]) et donc à considérer le fragment de poterie comme un document, l'important est ici la différence entre l'objet et le réseau qu'il forme avec d'autres objets, soit ce que Michel Foucault ([1977] 2001c) nomme un « dispositif ».

[23] On peut penser notamment à l'organisation hiérarchisée et l'indexation unique des fichiers organisés selon la métaphore du bureau, ainsi qu'à la conception de nombreux logiciels de manière à ce que le produit final demeure imprimable (Landow 1992 ; Nelson 2014).

[24] Le risque de confusion ici décrit n'est pas juste théorique et David M. Berry (2011b, 10) propose ainsi la « notion d'un *super-médium* », le code informatique, qui « n'est pas un médium qui *contient* les autres médiums, mais plutôt […] un médium qui les reforme et les transforme en une nouvelle forme unitaire » (nous traduisons). À supposer que le code confère une spécificité au contenu, on pourrait alors attendre qu'il altère significativement l'expérience sensible de ce dernier. Or, à prendre le cas de la production cinématographique, un spectateur lambda ignorant l'équipement de la salle de cinéma où il se trouve aura sans doute des

Cette spécificité des données numériques n'est donc pas liée à leur contenu mais à leur *paratexte* : « Parler de document *numérique*, c'est insister sur la facture technique, au détriment des autres caractéristiques » (Cotte 2004, 36). À condition de considérer avec Roland Barthes ([1973]) qu'un texte est une « pratique signifiante » sujette à interprétation et donc que le concept s'applique plus largement que sa simple déclinaison écrite, on peut en effet reprendre ici cette catégorie d'analyse de « paratexte » proposée par Gérard Genette (1987, 7) pour décrire l'ensemble des dispositifs qui effectuent une médiation entre le livre et son lectorat, « ce par quoi le texte se fait livre et se propose comme tel à ses lecteurs[25] ». Nombreuses sont les tentatives d'aborder les supports numériques, du livre au DVD, qui recourent au concept de paratexte (e.g. Morelli 2005 ; Vitali-Rosati 2015), et l'on ne fera donc ici qu'élargir cette approche à l'ensemble des formes physiques ou numériques que peuvent prendre un « contenant de données », c'est-à-dire dans le cas présent un fichier numérique[26].

difficultés à déterminer si la copie d'un film tourné sur pellicule et projeté sur écran géant est elle-même sur pellicule ou bien numérisée. Le code n'a d'ailleurs pas plus d'impact sur la capacité à analyser automatiquement un contenu : les capacités d'analyse automatisée de texte, au demeurant très limitées, sont ainsi difficilement transposables au médium cinématographique, pour lequel l'unité à partir de laquelle construire l'analyse de manière procédurale est plus difficile à choisir : le plan, l'image ou la seconde présenteront ainsi des limites empêchant de les considérer satisfaisant·e·s, la spécificité du médium cinématographique étant justement d'avoir reformé, transformé et unifié des médiums préexistants (enregistrement sonore, théâtre, cirque, photographie, etc.) en une nouvelle forme spécifique. De façon plus générale, un contenu ne saurait être segmenté sans être altéré : on n'analyse pas une suite de mots, de plans, de mouvements, etc., mais tout ou partie d'un ensemble, considéré de manière cohérente.

[25] La différence majeure entre Barthes ([1968] 1984) et Genette (1987, 8-9) tient dans la place de l'intentionnalité auctoriale, qui est centrale chez le second tandis que le premier se concentre sur la réception des textes. On élargit donc quelque peu l'approche de Genette (1987, 373), même si lui-même envisage la possibilité de penser un « paratexte hors littérature ». Par ailleurs, si le théoricien de la littérature français crée plusieurs subdivisions, notamment selon que le paratexte est intérieur (*péritexte*) ou extérieur au texte (*épitexte*) et selon qu'il est attribuable à une intention *auctoriale* ou *éditoriale* (Genette 1987, 10), il nous parait plus précis de créer une nouvelle catégorie paratextuelle spécifique au *support textuel*, qu'il soit physique ou numérique (livre broché, relié, de poche, numérique, etc.). Genette (1987, 20) ignore en effet le support lorsqu'il propose de distinguer entre un « péritexte » matériellement situé « *dans* l'espace même du volume » et un « épitexte » situé, « au moins à l'origine, *à l'extérieur* du livre » (nous soulignons). Il ignore en particulier les traductions et les rééditions posthumes, si ce n'est pour les comparer à l'édition originale, la seule qui l'intéresse vraiment au regard de l'intentionnalité, et évite notamment de parler de la publication en feuilleton, se restreignant donc quant à la question du support (Genette 1987, 11-12, 372-373).

[26] Il serait plus exact de parler ici de « fichier ou ensembles de fichiers » dans la mesure où un texte téléchargeable sur Gallica sera par exemple divisé différemment selon le mode de lecture, avec un fichier image par page lors de la lecture en ligne, mais un seul fichier texte ou fichier PDF par ouvrage au téléchargement, à moins de choisir l'option de télécharger une sélection de pages.

2.2 Données massives, modèle DIKW et simulation

Ce n'est ainsi que *de manière métonymique* que l'on peut confondre le livre ou tout autre objet et le fichier numérique qui en *simule* le contenu, quand bien même les métadonnées d'un fichier correspondront au livre original et non au fichier lui-même (cf. Cotte 2004, 40-41). À croiser les pensées de Genette et McLuhan, on peut en effet noter que les métadonnées sont formellement le paratexte du fichier et non du livre, médium dans le médium. Et si certains ensembles de données sont créés spécifiquement pour le numérique, comme Google Street View pour lequel aucun répertoire photographique d'une ampleur comparable ne pré-existait, ceci ne change toutefois rien au fait qu'un « objet » matériel – en l'occurrence la plupart du temps une rue ou une route et son environnement immédiat – se trouve simulé comme s'il s'agissait d'un décalque exact de la réalité. Or, on peut relever avec Jean Baudrillard (1981, 16) que la simulation « s'oppose à la représentation » justement en ce que la représentation suppose une reconnaissance de sa propre facticité, tandis que la simulation a vocation à se confondre avec le référent et donc à prétendre accéder à la réalité. Dans le cas de GSV, cette simulation s'avère plus évidente encore que dans le cas des livres numérisés puisque se trouvent simulées via l'interface à la fois l'immersion depuis un point fixe et la navigation entre plusieurs de ces points[27].

Ce constat d'une disjonction entre l'objet simulé et le fichier contenant la simulation n'est pas sans conséquence épistémologique, dès lors qu'on en revient à deux notions-clés pour l'étude des DM, à savoir celle de « donnée » et celle d'« information ». L'exploration de données (ED, *data mining*) est en effet le processus central dans toutes les approches automatisées des DM, puisque, dans ce cadre de pensée, « les données ne sont pas utiles en elles-

[27] Chaque photo panoramique de la base de données de GSV se trouve ainsi complétée automatiquement pour simuler un angle solide de 4π stéradians (celui d'une sphère), effaçant de fait du résultat final – sauf par des reflets accidentels – la voiture équipée pour la prise de vue et donc le principal outil de production de GSV. S'y ajoute la simulation 3D des façades de bâtiments, réalisée à partir de données géolocalisées captées en même temps que les images, qui permet au curseur de zoom de s'ajuster à la distance calculée à partir du point d'observation simulé. Cette mesure 3D permet de plus de simuler la navigation, avec la possibilité de cliquer sur un point de l'image correspondant à la route pour charger le panorama correspondant aux coordonnées de géolocalisation les plus proches du point visé (Anguelov et al. 2010). Enfin, l'effet d'immersion dynamique est rendu possible par une animation qui « simul[e] […] un fort sentiment de parallaxe » lors de la transition d'une « bulle » géolocalisée à une autre, en en dissimulant le caractère « discret », i.e. discontinu, de la simulation (Kopf et al. 2010, nous traduisons).

mêmes et d'elles-mêmes » mais « ont une utilité seulement si une signification et une valeur peuvent en être extraites » (Kitchin 2014b, 100). Dans une perspective ignorant l'évolution et la révision constante des connaissances scientifiques (cf. Latour 2008), le processus de production scientifique se trouve ainsi réduit à une « pyramide » à quatre étages où « les données précèdent l'information, qui précèdent la connaissance, qui précède la compréhension et la sagesse » (Kitchin 2014b, 9). Ce modèle DIKW (*data-information-knowledge-wisdom*), très repris dans la littérature consacrée aux DM, suppose de considérer les données comme une représentation symbolique des propriétés observables d'un objet, d'un phénomène ou d'un environnement, et de les distinguer de la notion d'information de manière « fonctionnelle, [et] non structurelle » (Frické 2009, 133, nous traduisons), le modèle étant pensé par Russel L. Ackoff ([1974] 1997) à des fins d'automatisation de l'analyse. L'information est ainsi souvent vue comme un produit manufacturé, c'est-à-dire en termes de données transformées (Rowley 2007, 171).

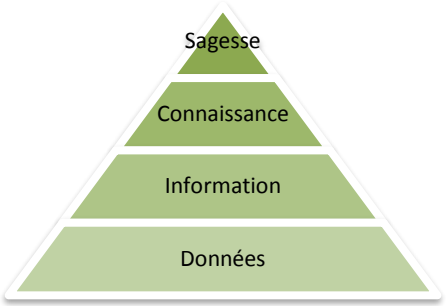


Fig. 1. Pyramide DIKW

Or, la faiblesse structurelle de ce modèle tient justement dans son incapacité à prendre en compte le caractère simulé des DM. Placées à la base de la pyramide (cf. fig. 1), les données ne peuvent en effet pas permettre d'accéder à l'ensemble des propriétés de l'objet observé, la simulation entrainant une nécessaire déperdition qui se répercute ensuite à chaque stade supplémentaire de l'analyse automatisée[28]. Au deuxième niveau, l'extraction de l'information ne peut non plus être exhaustive face à l'infinité des relations de corrélation théoriquement identifiables, et ce, même à envisager des algorithmes autogénérés à la chaîne par apprentissage machinique (Simonite 2017a ; Zoph et Le 2017) : on ne saurait par exemple envisager de répliquer automatiquement

[28] Cette incapacité est particulièrement évidente dans des cas de réutilisations de données collectées auparavant avec d'autres objectifs, comme dans le cas des « dossiers passager » des compagnies aériennes (PNR, *personal name record*), croisés après le 11-Septembre avec des listes de personnes supposées dangereuses. Ce détournement d'usage n'a ainsi pas manqué de créer de faux positifs menant à des expulsions du territoire pour crime d'homonymie (GAO 2005). Il a aussi favorisé, pour contourner le fait qu'il n'avait pas été prévu à cette fin, le développement d'un profilage islamophobe et/ou raciste, par inférence à partir de données comme le nom et le prénom ou la consommation d'un repas halal à bord du vol (House of Lords 2007, 15).

une enquête origine-destination (OD) sans l'établissement d'une stratégie de recherche spécifique ni sans une réflexion éthique sur la source et les modalités d'utilisation des données. Enfin, la montée en généralité suivant une analyse automatisée des informations suppose de s'éloigner d'une science basée uniquement sur les données pout raisonner par inférence[29] (Frické 2009, 135). Cette ignorance de nombreuses questions épistémologiques nous encourage enfin à remplacer la « sagesse » par l'« hubris » au sommet de la pyramide, celle-ci demeurant de toute façon rarement définie et encore moins opérationnalisée (Rowley 2007, 178).

2.3 Les données comme outil dans le cadre d'une stratégie de recherche

Le concept de donnée peut tout d'abord être défini, en science de l'information, à travers un « triplet < *e*, *a*, *v* >, où *e* est une entité dans un modèle conceptuel, *a* est un attribut de l'entité *e*, et *v* est une valeur du domaine de l'attribut *a* », et où de plus « l'entité *e* a une valeur *v* pour l'attribut *a* ». Par exemple, dans un système de ressources humaines, « < John Doe, département, droit > » indiquera que l'employé John Doe travaille au département de droit (Redman, Fox et Levitin 2003, 784, nous traduisons). Dans cette première approche, les auteurs mettent, à raison, de l'avant la dimension de « description abstraite » de la modélisation, et donc son *incomplétude*, qui a pour contrepartie de pouvoir décrire de manière cohérente tout autant des entités abstraites que des entités concrètes.

Cette première définition se distingue donc d'ores et déjà d'une définition « *computationnelle* » des données assimilant le fichier à son contenu abstrait et considérant en conséquence les données comme des « ensembles […] d'*éléments binaires* […] traités et transmis électroniquement par des technologies comme les ordinateurs et les téléphones portables » (Floridi 2008, 235, nous traduisons). Elle permet également de pointer la limite d'une définition « *informationnelle* » des données, qui correspond, dans une démarche d'ED, à traiter les données

[29] Par exemple, dans le cas de Google Flu Trend, il n'est plus question de proposer une courbe basée sur les données de grippe dès lors qu'une estimation est formulée à partir des données de recherche Google. Il est en effet supposé que la corrélation entre l'évolution du nombre de certaines requêtes et celle du nombre de cas de grippe, établie dans la phase de préparation, va se maintenir dans l'avenir.

comme de l'information en puissance (Floridi 2008, 234) et qui ne diffère de la définition computationnelle que par l'étage de la pyramide DKIW à laquelle elle se situe : par exemple, considérer les données produites par Walmart avec comme unité de mesure l'objet « bibliothèque du Congrès » (cf. 1.3) suppose que les données ne sont pas médiées par des signes mais matériellement inscrites sous forme binaire sur un disque dur, un serveur informatique, etc.

Ce rassemblement définitionnel nécessite de suivre ici l'approche de Ferdinand de Saussure ([1916] 1995, 97-99), selon qui le « signe linguistique unit non une chose et un nom, mais un concept et une image acoustique », cette dernière étant définie comme « l'empreinte psychique » d'un « son matériel, chose purement physique », « la représentation que nous en donne le témoignage de nos sens ». Le concept de signe – dont on étend ici avec Barthes (1964) le domaine d'application – est donc abstrait et analytiquement distinct de son vecteur matériel, en réponse au constat que la langue n'est pas une « nomenclature, c'est-à-dire une liste de termes correspondant à autant de choses », puisque le concept (« *signifié* ») comme sa représentation (« *signifiant* ») peuvent être soumis à interprétation. Autrement dit, Floridi (2008, 234) distingue ces deux approches parce que lui-même en accepte les prémisses, comme en témoigne sa critique d'un défaut de prise en compte des enjeux de la « *compression des données* » ou de leur « *cryptographie* » adressée aux tenants d'une définition qualitative – qualifiée plutôt d'« *épistémique* » au sens anglo-saxon du mot (*knowledge-oriented*).

Dans cette troisième approche qualitative, les données sont ainsi comprises comme un « ensemble de *faits* » susceptibles de servir de « base à un raisonnement », et le reproche formulé par le philosophe italien est surtout de ne faire que déplacer le problème de la définition du concept de donnée vers un nouvel élément atomique, le « fait ». Lui-même propose dans ce cadre une quatrième définition, qu'il qualifie de « diaphorique », laquelle vise à réconcilier les différentes approches en décrivant ce qui distingue une donnée d'une autre par « un manque d'uniformité » : « (D) donnée = x différent de y, où x et y sont deux variables non interprétées » (Floridi 2008, 235). Bien qu'heuristiquement productive, cette définition de-

meure relative à chaque projet de recherche[30] et ignore la nécessité du signe. L'exemple extrême cité par le philosophe, celui d'une page blanche n'est en effet un signe que s'il est *interprété* comme tel, et ne pourra donc constituer une « variabl[e] non interprété[e] », à la différence du symbole « ∅ » introduit par le groupe Bourbaki à la fin précise de désigner un ensemble nul (Miller [c2008] 2017).

Pour Floridi, la question semble devoir être d'établir un critère de distinction entre les données, considérées comme un élément « insécable » analytiquement. Or, à reprendre cette perspective « diaphorique » en rapport avec la définition qualitative des données – soit la troisième dans le système de Floridi –, rien n'oblige à objectiver une telle distinction dès lors qu'on assume la dimension abstraite du concept de « donnée », et qu'on le considère comme un outil de représentation au service d'une stratégie de recherche. Sans entrer dans les détails du débat sur les contenus conceptuels de la perception (cf. Bricka 2016), on peut en effet estimer, avec Maurice Merleau-Ponty ([1945] 2008, 32-34) que « nous supposons d'emblée dans notre conscience des choses ce que nous savons être dans les choses ». Autrement dit, un contenu ne sera jamais complètement perceptuel et l'acceptation de l'existence d'un « réel » ne nous le rend pas pour autant accessible dans son intégralité. La « qualité déterminée » d'un « phénomène positif », « par laquelle l'empirisme voudrait définir la sensation » se trouve donc être « un objet, non un élément de la conscience », et ne saurait à ce titre être considérée objective.

Face à ce que Wilfrid Sellers ([1956] 2000) appelle « le mythe du donné (*given*) » (nous traduisons), on peut donc considérer le concept de « donnée » comme un outil mobilisé dans le cadre d'une stratégie de recherche visant a) à contourner l'inaccessibilité du réel à la perception objective et intégrale, et b) à rationaliser les contenus perçus par la médiation d'une grille conceptuelle afin de pouvoir en extraire, par un processus d'analyse et de synthèse, des

[30] Parmi les différentes distinctions possibles, Floridi cite notamment la différence « *de dicto* », avec une distinction entre deux « symboles d'un code » comme « les lettres A et B dans l'alphabet latin ». Mais on peut tout aussi bien vouloir distinguer, selon les besoins, les lettres « A » et « B » que « A » et « a », « a » manuscrit et « a » imprimé, etc.

informations pertinentes à une ou plusieurs questions de recherche. La différence entre l'approche « épistémique » (qualitative) et l'approche « computationnelle » des données – dans laquelle on inclut l'approche « informationnelle » – semble alors être que la première tend vers la représentation du réel tandis que la seconde tend vers sa simulation, sans bien sûr que cette différence puisse pour autant être considérée comme aussi significativement marquée dans la pratique puisque la confusion entre des indicateur et le « réel » ne nécessite pas que ceux-ci se trouvent quantifiés. Floridi nous semble en résumé avoir raison quant à la possibilité de formuler une définition applicable indifféremment aux domaines qualitatif et quantitatif, à la condition d'ajouter à son approche « diaphorique » un « critère de modestie », c'est-à-dire la conscience que les données constituent une représentation et non un accès à la réalité.

2.4 L'information, notion contaminée

Cette définition « tactique » (de Certeau [1980] 1990, 51) prend donc en compte l'incomplétude structurelle de l'information synthétisée à partir des données et renforce l'obligation de rigueur dans l'analyse, en particulier en ce qui a trait à la généralisation des conclusions. Or, la confusion entre support et contenu dans l'approche de la notion de donnée se répercute nécessairement à l'étage suivant du modèle DIKW, celui de l'information : si l'information est considérée comme constituée de données « raffinées » – ou autre métaphore extractiviste –, alors cette information demeure, tout autant que les données, considérée comme constituée d'octets plutôt que de signes porteurs de sens.

La contagion évoquée en 1.1 est d'ailleurs loin d'être unidirectionnelle, puisqu'une stricte distinction entre une définition de l'information par le sens et une autre par le volume est explicitement défendue par Claude E. Shannon ([1948] 1963, 31) dans un texte fondateur de la théorie de l'information où le mathématicien et ingénieur étasunien explique que les « aspects sémantiques de la communication ne sont pas pertinents au problème d'ingénierie » qui l'occupe (nous traduisons). Or, si la définition de Shannon est différente de celle du modèle

DIKW[31], c'est bien lui qui a importé ce terme en informatique, envisageant même de le changer pour celui de « communication » face aux mésinterprétations répétées de son propos. La notion, lui ayant échappé, sera toutefois reprise de manière de plus en plus distanciée de son sens premier, jusqu'à « être appliqué[e] à tout ce qui pouvait être métaphoriquement interprétée comme un "message[32]" », au détriment des « distinctions intellectuelles » (Roszak 1986, 12-14, nous traduisons).

Face à ce constat, Theodore Roszak (1986, 96-97, 105) en vient même à remettre en question la pertinence de la notion d'« information », en estimant que « les idées créent les informations, pas l'inverse » et qu'il n'est pas nécessaire d'accepter la fiction selon laquelle l'informatique répliquerait le fonctionnement du cerveau humain. On a vu en effet que les données n'étaient jamais brutes : comme l'expliquent Lisa Gitelman et Virginia Jackson (2013, 3), « [l]es données ont besoin d'être imaginées *comme* données » avant d'exister, ce qui nécessite toujours interprétation (nous traduisons). De ce fait, la stratégie de constitution des données est d'ores et déjà porteuse de sens, et détermine d'avance quels types d'information pourront être retenus pour l'analyse. L'océrisation de textes, par exemple, ne fait pas juste accumuler des données textuelles mais les accumule *à des fins* de traitement automatisé du langage naturel (TAL, *natural language processing*). Considérer l'information sémantique

[31] Comme l'explique Warren Weaver ([1949] 1963, 8-9), la théorie de l'information attribue au mot un « sens spécial » au contexte computationnel, qui « ne doit pas être confondu avec [celui de] signification (*meaning*) » et qui lui est même directement opposé (nous traduisons). L'information est dans ce cadre définie par une « quantité de liberté de choix », selon une formule qui correspond « dans les cas les plus simples » au « logarithme du nombre de choix disponibles », spécifiquement un logarithme de base 2 dans une situation de binarité (« 0 » ou « 1 ») comme avec les bits. À reprendre la pyramide DIKW, l'information, dans son sens informatique, précède alors en fait la donnée.

[32] Hans-Jörg Rheinberger ([2006] 2010, 208-211) insiste par exemple sur le caractère contingent de l'émergence du concept d'« information » en génétique via les travaux de François Jacob, d'abord en 1960 comme strict synonyme de « spécificité structurelle » pour progressivement désigner un « message », inscrit dans l'ADN et « traduit » par un « programme » (nous traduisons). En 1965, le généticien français finit même par parler de « texte chimique » et d'« écriture de l'hérédité ». L'inspiration, à la fois informatique et linguistique, mène donc à une métaphore simplificatrice qui tend à faussement objectiver l'information et même à essentialiser la notion puisqu'un enchaînement de bases nucléiques (adénine, cytosine, guanine ou thymine) sera censé contenir une information objective, créant une stricte équivalence entre le signifiant et le signifié. En retour, cette essentialisation de la métaphore de la notion d'information génétique pourra être réimportée en SHS, par exemple par Alberto Piazza (2005, 95-96) comme argument de légitimation de l'approche de Franco Moretti (cf. 3.3).

comme le résultat de l'organisation de données (Feather 2013, 474) renvoie donc à une fiction qui distingue les données du sens qu'on peut leur associer à des fins techniques plutôt qu'analytiques, et placer le curseur de l'analyse sur la mise en relation des données ne signifie pas que les données elles-mêmes ne soient pas porteuses de sens[33]. Pour toutes ces limites, la fiction définitionnelle informatique n'a pas de raison d'être transposée à la recherche en SHS : les limites actuelles de l'analyse automatisée constituent en effet un problème à dépasser et non un horizon souhaitable pour l'ensemble des approches.

À toutes fins pratiques, on distinguera ainsi la collecte de données comme une étape parmi d'autres d'un processus analytique à la fois itératif et incrémental, à l'image par exemple du modèle proposé par Matthew B. Miles et Michael Huberman ([1983] 1994, 10-12) (cf. fig. 2). Plutôt que de prétendre extraire la substantifique moelle des données, on prend ainsi en compte l'informalité d'une part importante du processus de production de la recherche, c'est-à-dire la part de savoir-faire qui se prête mal à une transcription écrite selon les normes scientifiques en vigueur (Becker [1986] 2007 ; Delamont et Atkinson 2001).

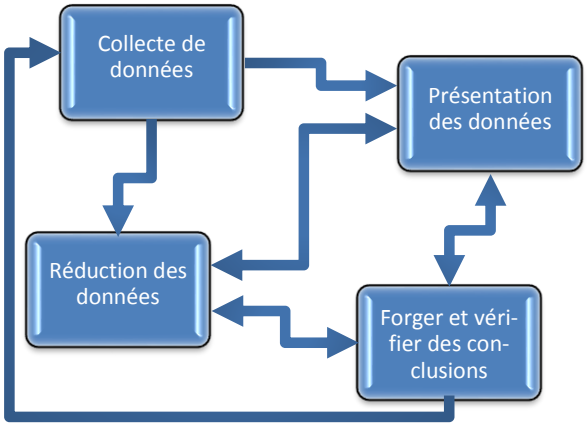


Fig. 2. « Composantes de l'analyse de données : modèle interactif » (adapté et traduit de Miles et Huberman [1983] 1994, 12)

3. La lecture intertextuelle comme réponse qualitative aux prétentions hégémoniques des humanités numériques

On a donc vu que l'expression « données massives » désigne en fait des données qualitatives sous forme numérisée, et que leur appréhension informatique impliquait de penser dans le cadre d'une simulation. On a en conséquence montré la nécessité de distinguer les concepts

[33] La formule « information = donnée + sens », récurrente dans la littérature (e.g. Napoli 2005, 2), constitue ainsi un slogan et non une équation, puisque cette dernière supposerait sinon que « information - donnée = sens », c'est-à-dire que le sens peut exister indépendamment des données elles-mêmes.

développés à des fins computationnelles de ceux pertinents à l'analyse scientifique, en nous arrêtant sur l'intérêt d'appréhender la notion de « donnée » relativement à la stratégie de recherche dans laquelle elle s'inscrit et en pointant la non-pertinence à reprendre à des fins épistémologiques les distinctions d'une pyramide DIKW construite en fonction des limites techniques de l'analyse automatisée. On introduira à présent à l'intérêt d'une analyse distincte de cette fiction informatique et davantage centrée sur la représentation, à travers une critique des prétentions hégémoniques des humanités numériques (HN) et la proposition d'un cadre de lecture alternatif, vers une approche intertextuelle des DM.

3.1 Humanités numériques et fétichisme de la technologie

Nombreux ont été les manifestes publiés au cours de la dernière décennie en défense des humanités numériques (HN), témoignant d'un enthousiasme croissant pour les perspectives offertes par les progrès de l'apprentissage machinique (AM) et plus spécifiquement de l'apprentissage profond[34] (AP) (Schnapp et al. 2009 ; Dacos 2011 ; Arnaud et al. 2013). Cette centralité des outils et de leurs potentialités doit d'ailleurs être considérée fondatrice des HN, Jeffrey Schnapp et al. (2009, 2) définissant par exemple leur discipline davantage par les outils utilisés que par un projet épistémologique commun :

> Les humanités numériques ne sont pas un champ unifié mais *un ensemble de pratiques convergentes* qui explorent un univers dans lequel : a) l'imprimé n'est plus le medium exclusif ou normatif par lequel la connaissance est produite; et b) les outils, techniques et médias numériques ont altéré la production et la dissémination de la connaissance dans les disciplines artistiques et dans les sciences humaines et sociales (nous traduisons).

[34] On n'entrera pas dans l'histoire des HN, auxquelles on trouve généralement et rétrospectivement comme précurseurs la littérature et la linguistique computationnelles (LLC, *literary and linguistic computing*) puis les humanités computationnelles (HC, *humanities computing*) (Burnard [2009] 2012). Une « contre-histoire » des approches qualitatives du numérique demeurerait par ailleurs à écrire, avec probablement comme point de départ le memex de Vannevar Bush (1945), et comme moment charnière l'émergence d'une théorie de l'intertextualité (cf. Landow 1992).

Si cet aspect n'est pas toujours évident au premier abord, tant l'accent est mis sur la diversité des possibilités ouvertes par les évolutions de l'environnement numérique, le principal élément définitionnel des HN est à ce titre *l'exclusion* de toute possibilité d'un « virage numérique qui pourrait d'une quelconque manière laisser intactes les sciences humaines » (Schnapp et al. 2009, 13). En clair, la possibilité que des méthodes préexistantes, notamment qualitatives, puissent se trouver applicables à l'environnement numérique semble inconcevable aux prophètes d'une nouvelle épistémologie centrée sur le tout-automatisé. C'est également ce qu'exprime Marin Dacos (2011) lorsqu'il expose dans son manifeste que « [l]es *digital humanities* désignent (sic) une transdiscipline, *porteuse des méthodes, des dispositifs et des perspectives heuristiques liés au numérique* dans le domaine des sciences humaines et sociales » (nous soulignons).

Cette affirmation répétée contraste avec celle de la nécessité d'une « communauté de pratique solidaire, ouverte, accueillante et libre d'accès », puisqu'on voit mal comment les HN pourraient prétendre s'appuyer « sur l'ensemble des paradigmes, savoir-faire et connaissances propres à ces disciplines » en se limitant aux seuls outils et perspectives spécifiques au « champ du numérique ». Le corollaire de cette exclusion est en effet le fait que les HN reposent sur l'application d'une méthode *computationnelle* à des enjeux de SHS (Meeks 2014). Un outil logiciel sera donc nécessairement mobilisé *à des fins d'analyse automatisée*, qu'il s'agisse par exemple d'un système d'information géographique (SIG, *geographic information system*) ou d'un logiciel de traitement automatisé de la langue naturelle (TAL).

Les raisons de cette invisibilisation des approches non-automatisées peut sembler obscure, puisque nombreuses sont les tâche d'interprétation non-réplicables par un biais automatisé, de l'analyse filmique à l'analyse d'entretiens. Les « changements dans le langage, les pratiques, les méthodes et les résultats » défendus par Schnapp et al. (2009, 3) ne sont ainsi pas censées signifier de faire « table rase du passé », à en croire Dacos (2011), qui parle plutôt de s'appuyer sur « l'ensemble des paradigmes, savoir-faire et connaissances propres » aux disciplines. Mais leur prise en compte n'a de valeur en HN que lorsqu'elles sont utilisées au travers des « outils et [d]es perspectives singulières du champ numérique », autrement dit à

travers une analyse automatisée. Face aux « promesses hors-sol » des HN (Bouchet, Carnino et Jarrige 2016) et face à la préférence pour l'outil au détriment de la méthode, on voit donc mal en quoi il serait légitime de reprocher au « monde universitaire » de ne pas « évolu[er] au même rythme » (Arnaud et al. 2013).

On propose toutefois de comprendre ce rejet par l'hypothèse d'un rapport fétichiste à la technologie, où le caractère automatisé de l'analyse serait censé suffire à lui conférer une objectivité. Le fétichisme, dans la théorie marxienne, est en effet défini par le fait de conférer une caractéristique sociale (ici, l'objectivité dans l'analyse) à un objet qui ne saurait en posséder[35] (Marx [1867] 1993, I, 1, 4). Les résultats d'une analyse automatisée se voient en conséquence autonomisés, « dou[és] d'une vie propre, entretenant des rapports les uns avec les autres et avec les humains ». À l'image de la « forme-monnaie » pour la marchandise, la forme numérique joue pour la connaissance scientifique un rôle d'« occult[er] sous une espèce matérielle, au lieu de le révéler, le caractère social » de la production. De ce fait et comme le résume David Harvey (2003, 3), la technologie se voit investie, par « une croyance naïve », du pouvoir de « fournir une solution à n'importe quel problème » (nous traduisons). Plus précisément, une confusion entre support et méthode s'ajoute à la confusion entre contenu et support déjà décrite en 2.1, et la neutralité prêtée à la forme numérique des données se « transmet » censément à l'analyse automatisée, du fait de sa forme également numérique[36].

L'hypothèse d'un fétichisme de la technologie permet de résoudre le problème du seuil d'automatisation à partir duquel il est possible de prétendre se réclamer des HN, soulevé par

[35] Elle se distingue donc d'un « quiproquo » qui transfère simplement des propriétés « physique[s] », comme celui qui mène à confondre « l'excitation du nerf optique proprement dit » avec la « forme objective d'une chose à l'extérieur de l'œil » (Marx [1867] 1993, I, 1, 4).

[36] Les espoirs investis dans les possibilités de l'apprentissage machinique mènent même à anthropomorphiser les algorithmes en leur prêtant une volonté et une intelligence propres, avec toujours en sous-texte un imaginaire de science-fiction. L'incapacité d'un programme à prendre en compte la casse des caractères pourra par exemple être considérée comme le signe de la nécessité de mieux « parler le langage de l'ordinateur » (Wilson 2017, nous traduisons). Cette tendance est décelable dans le discours journalistique, mais également dans le discours ingénieur, à parler par exemple de « rêves » pour désigner le résultat du renversement du fonctionnement d'un réseau neuronal à des fins de génération graphique (Mordvintsev, Olah et Tyka 2015, nous traduisons). Le vocabulaire informatique repose d'ailleurs de façon générale sur l'anthropomorphisation des algorithmes, avec des termes et expressions comme « apprentissage », « incitatif », « intelligence », « réseau neuronal », etc.

Marin Dacos et Pierre Mounier (2015, 7) face au constat que « les technologies numériques […] sont désormais quasiment universelles » et que « tout chercheur utilise ne serait-ce qu'un outil bureautique au cours de sa journée de travail » :

> Fait-on des humanités numériques lorsque l'on rédige son article sur un traitement de texte ou que l'on échange par e-mail avec un collègue ? Ou bien faut-il réserver ce terme à des usages beaucoup plus sophistiqués, lorsque les sources de la recherche sont traitées au moyen de bases de données par exemple ?

En effet, au-delà sa familiarité d'usage, un logiciel comme Microsoft Word fait preuve d'une grande complexité dans ses fonctions de traitement de texte, qu'il s'agisse de la saisie de caractères, de la mise en forme du texte, de la révision orthographique et grammaticale, de propositions de synonymes, etc., le logiciel faisant même désormais appel à l'apprentissage machinique dans le processus de révision. La métaphore du « chapiteau » (*big tent*), censée exprimer une « approche œcuménique » en réponse à cette contradiction (Terra 2011, nous traduisons) ne peut ainsi être considérée adéquate, puisqu'on a vu que le recours à des méthodes automatisée constituait de fait une condition d'intégration aux HN. Gregor Wiedemann (2013, 334), partisan d'un rapprochement entre approche qualitative et HN, note lui-même la distance entre cette dernière approche et l'analyse qualitative de données (AQD, *qualitative data analysis*), qui consiste, sur des logiciels comme NVivo, en une réplique numérisée de tâches de « codage et prise de notes auparavant réalisées avec des stylos et des surligneurs, des ciseaux et de la colle » (nous traduisons). Même à faire de l'analyse lexicométrique, comme le permet par exemple le logiciel Alceste (Roy et Garon 2013, 155; Khelifi 2017), celle-ci demeurera un préalable à l'interprétation qualitative, laquelle n'en deviendra pas pourtant autant automatisée[37].

[37] Cette affirmation demeure assez peu contestée, malgré une confusion entretenue, notamment par Schnapp et al. (2009, 2), sur l'entrée des HN dans une phase de « nature *qualitative, interprétative, expérientielle, émotionnelle, générative* ». Leighton Evans et Sian Rees (2012, 21, 30) minimisent ainsi « l'importance de l'interprétation herméneutique humaine », mais sans en nier la spécificité non réplicable : « l'exploration de données (*text mining*) peut aider l'analyse herméneutique – pas la remplacer » (nous traduisons).

Autrement dit, c'est à la condition que l'*analyse* des données qualitatives se limite à une quantification, sans interprétation subséquente, qu'on peut parler d'HN, et non directement parce que le support des données est numérique, et toute approche proprement qualitative, prenant en compte la nécessité d'une interprétation, se trouve de fait hors du cercle des HN. À prendre une définition très procédurale, « [l]'analyse qualitative peut être résumée, dans son essence, à un passage en revue attentif et répété des données, une catégorisation, une interprétation et une rédaction » (Schönfelder 2011, nous traduisons). Or, si les deux premières tâches peuvent être assistées par un logiciel, les deux dernières, soit l'interprétation et sa mise en forme sont encore impossibles à répliquer de manière automatisée[38]. Les parenthèses sont donc trompeuses lorsque Gregor Wiedemann (2016, 2) prône une « analyse de texte (semi-)automatique assistée par ordinateur » (nous traduisons) : de la même manière qu'une tournée de distribution de courrier ne saurait être qualifiée de « semi-motorisée », l'interprétation d'un texte n'est pas plus « semi-automatique » que le reste des tâches de recherche, aujourd'hui quasi intégralement assistées par ordinateur.

3.2 Tournant computationnel contre prétention hégémonique

Cette définition des HN par la primauté de l'outil technologique et de la datafication (cf. 2.1) suppose qu'elle se fond sur les présupposés d'un fétichisme technologique, qui mène à ignorer certaines limites méthodologiques et techniques de la discipline. En lieu et place d'une autoréflexivité critique, s'impose ainsi l'argument qu'il « serait inepte » de « se passer des données numériques » et que « les technologies numériques […] nous permett[ent] de faire des choses de façon plus simple, plus efficace[39] » (Burnard [2009] 2012). L'efficacité étant

[38] Plus exactement, l'automatisation de l'écriture n'est possible que de manière extrêmement limitée, par exemple pour compléter un texte à trous avec des données comme la magnitude ou le lieu d'un séisme mises en ligne de manière standardisée par une autorité publique (Oremus 2014). Et encore une tâche aussi basique n'empêche-t-elle pas les erreurs, comme en 2017 avec une fausse alerte suite à une correction portant sur les coordonnées d'un séisme datant de 1925 (Lin 2017).

[39] La comparaison avec les méthodes quantitatives est mise sous le tapis par leur association implicite au support de « l'édition papier », renvoyé à une « époque » aujourd'hui « révolue » (Burnard [2009] 2012). La fin du papier, prophétisée sans discontinuer depuis au moins la fin des années 1990, rencontre pourtant toujours des obstacles majeurs à sa réalisation, notamment au niveau du coût et de la capacité de partage des fichiers (Kovač 2008). La cohabitation objective entre les différents supports n'a donc pas de raison de disparaître à court terme.

surtout définie par les critères et les objectifs contingents qu'on lui associe (Carlier 2015), on peut voir dans l'insistance à se présenter comme en possession d'une réponse méthodologique unique l'expression d'une prétention hégémonique, rationnalisée par le biais du fétichisme de la technologie.

Ce fétichisme de la technologie permet de comprendre que les HN constituent une tentative de transposition aux SHS d'une logique issue de l'informatique, soit une imitation quelque peu naïve de méthodes considérées « plus scientifiques » *parce que* médiées par la technologie[40]. Reconnaître l'existence d'un « tournant computationnel » où « la recherche est de plus en plus médiée par la technologie numérique » (Berry 2011a, 1, nous traduisons) n'implique en effet en rien d'avoir à abandonner toute approche qualitative (cf. 3.4), et l'appréhension de l'impact de cette évolution sur les « approches épistémologiques et ontologies qui sous-tendent un programme de recherche » (Berry 2011a, 1) s'avère d'autant plus complexe que l'omniprésence actuelle de l'outil numérique dans le processus de production de la recherche tend trop souvent à être confondue avec une *tabula rasa*, le fétichisme technologique allant de pair avec la limitation de la réflexion à une dimension paratextuelle (cf. 2.1).

Or, c'est justement sur la base de cette fausse spécificité des documents numérisés que David M. Berry (2011a, 2-4) estime que le caractère « *discret* » de l'encodage « transforme [le] flux continu de la réalité quotidienne [en] une grille de nombres qui peut ensuite être manipulée en utilisant des algorithmes ». Si une telle approche permet certainement de produire de « nouvelles connaissances » et de nouvelles « méthodes pour le contrôle de la réalité », elle n'en modifie pas pour autant ladite réalité, dont la compréhension demeure l'objectif et il est donc très contestable d'affirmer que « les changements médiaux produisent des changements épistémologiques » sans prendre en compte en fonction de quel média un contenu est produit.

[40] Lou Burnard ([2009] 2012) estime ainsi que « traiter le texte comme s'il s'agissait d'une donnée » participe à dépasser « une confrontation entre le texte et les données », là où elle ne fait de la sorte qu'importer en SHS une approche en réponse à des contraintes techniques spécifiques à l'informatique.

Du point de vue de ce « tournant », et sauf par une illusion d'objectivité issue de la conception computationnelle des données, les HN n'ont enfin aucune raison d'être considérées comme se situant davantage « à l'intersection de la science informatique et des sciences humaines » (Biemann et al. 2014, 80, nous traduisons), qu'une autre tendance des SHS intégrant à la recherche des méthodes et outils issu·e·s de l'analyse informatique. Or, cette prétention à pouvoir se réclamer de l'analyse informatique n'a pas lieu sans arrière-pensées, puisque l'élément distinctif fondamental des HN, dont l'exclusion des approches qualitatives s'avère le corollaire, s'avère l'objectif de créer en SHS un « paradigme computationnel » unifié, présenté par Dacos et Mounier (2015, 15) comme « la théorie de l'information » où « l'objet étudié est converti, manipulé, analysé sous une catégorie commune : l'information, objet de calculs ». De même Berry (2011a, 9) parle-t-il d'un changement de paradigme « dans le sens kuhnien ».

Mais une révolution scientifique suppose justement, pour Thomas Kuhn ([1962] 1996, 74-75), le constat d'une « crise », c'est-à-dire d'une inaptitude répétée et persistante des scientifiques à ajuster un modèle pour en combler les insuffisances (nous traduisons). En l'occurrence, le constat d'une telle crise des SHS serait en théorie aisé à établir, puisque le « tournant computationnel » date déjà d'il y a plusieurs décennies[41]. Or, la supposée nécessité de repenser « ce qui est humain dans les humanités *computationnelles* ou sciences sociales » avancée par Berry (2011b, 21) n'est que la conséquence de l'affirmation infondée selon laquelle le monde tendrait à être considéré de plus en plus de manière discrète plutôt que continue.

De plus, le problème à faire appel à Kuhn pour légitimer une approche de SHS tient dans le fait que le philosophe des sciences pense les différentes sciences comme un tout unifié subordonné au modèle des sciences naturelles et formelles (SNF), là où son approche n'y est

[41] En 1968 déjà, François Furet (1968) pouvait ainsi préparer pour *Le Nouvel Observateur* un dossier titré « Comment l'informatique bouleverse les sciences humaines ». Le fond de notre critique y était d'ailleurs déjà porté par Raymond Boudon (1968, 38), qui insistait sur la nécessité de ne pas attribuer « de mystérieuses et anthropomorphiques vertus » à un outil qui permet seulement la « rapidité » dans des « opérations logiques et arithmétiques ».

pas transposable rigoureusement : certain de la supériorité des SNF, il considère que l'émergence d'un paradigme unifié distingue ce qui constitue une science de ce qui n'en constitue pas une, et donc que les SHS n'ont pas encore effectué leur « transit[ion] vers la maturité » (Kuhn [1962] 1996, 21-22). Plutôt que de reconnaître qu'il échoue à transposer son modèle aux SHS, il prive ainsi ces dernières du titre de « science » et ouvre la voie à toutes les instrumentalisations de sa théorie, qu'il s'agisse de rejeter les SHS dans leur ensemble (Dépelteau [2000] 2003, 386-387) ou, comme ici, d'en hiérarchiser les approches en fonction d'une lecture imitative des SNF[42].

Dans ce contexte, plutôt que de prétention paradigmatique, il est plus exact de parler de prétention hégémonique, c'est-à-dire d'une aspiration à devenir une référence méthodologique normalisée (cf. Hall [1980] 1997, 69), en ce qui a trait au moins à la recherche en terrain numérique[43]. À la différence de ce qui serait nécessaire à fonder un paradigme, les HN ne permettent en effet pas de produire une interprétation du contenu des DM. La volonté d'imposer comme évidente et inévitable une position qui ne mérite pas de l'être, même partiellement excusée par un excès d'enthousiasme, nous semble donc suffisante pour qualifier d'hégémoniques les ambitions des HN, dans lesquelles le fétichisme de la technologie joue pleinement son rôle, puisque l'approche automatisée est censée à elle seule permettre de parvenir à une démarche objective et plus « efficace » que les autres.

3.3 « Lecture distance » et invisibilisation de l'approche qualitative

C'est d'ailleurs le principal reproche qu'on peut adresser aux HN que cette prétention hégémonique, plus ou moins consciente et assumée selon les auteur·e·s mais corollaire de la certitude de participer à un bouleversement épistémologique par ce qui n'est que la redécouverte

[42] Et encore le terme d'« imitation » est-il un peu trop positif, puisque l'objectif des recherches en intelligence artificielle n'est pas d'imiter le fonctionnement du cerveau mais d'en *répliquer* les performances en s'inspirant de son fonctionnement (Le Cun 2016).
[43] Berry (2011b, 21-22, 128) précise que l'émergence d'un « paradigme » computationnel ne signifie pas forcément « que les méthodes et pratiques *existantes* des sciences informatiques deviennent hégémoniques », mais de *nouvelles* méthodes et pratiques semblent toutefois appelées à le devenir selon lui, tandis que « l'ontologie du computationnel » devient « de plus en plus hégémonique ».

de l'eau tiède, soit en l'occurrence celle des données qualitatives[44]. En effet, on peut exprimer ses doutes face au miracle annoncé d'un bouleversement paradigmatique, a fortiori lorsque les DM ne sont que peu utilisées dans des travaux d'HN, en raison de ce que Gerhard Heyer et Marco Büchler (2010, 526) nomment des « problèmes d'infrastructure[45] » (nous traduisons). Parmi ceux-ci, on peut notamment retenir les différents formats des bases de données, la qualité variable des métadonnées[46] ou de l'océrisation[47], les options de recherche[48], ou les options et formats de téléchargement. S'y ajoute les questions du copyright, des formats propriétaires et du code source non ouvert, des formats en colonnes mal transcrits, des documents non océrisés, et ainsi de suite. Une standardisation semble des plus lointaines et rend donc plus fastidieuse l'analyse automatisée des données pour répliquer ce qu'une analyse « manuelle » permet sans difficulté.

Le principal outil disponible qui fasse appel aux DM, soit Google Ngram Viewer, est par ailleurs un exemple de ce à quoi les HN ne sont pas en mesure de prétendre, avec un fort contraste entre la prétention de Michel et al. (2011) à fonder une nouvelle discipline, la « culturonomie », et la pauvreté des conclusions proposées dans plusieurs exemples produits à

[44] La critique mérite certes quelques nuances, en particulier en référence aux arguments avancés en défense de l'utilisation de méthodes issues des HN *dans un champ disciplinaire précis*, qui nécessitent de reconnaître la spécificité de ce dernier et donc l'incapacité des HN à l'englober dans un nouvel ensemble. C'est le cas par exemple de Ted Underwood (2013, 170), qui voit dans l'introduction des HN en études littéraires une « saine diversification méthodologique », en contrepoids à une surreprésentation des études de cas dans la discipline (nous traduisons).

[45] Les projets d'ED existants sont ainsi souvent effectués sur des bases de données présentant une transcription de grande qualité, par exemple des corpus de textes antiques, et se trouvent en l'état difficilement transposables aux bases en ligne comme Google Books ou Gallica.

[46] Les erreurs de métadonnées les plus dommageables pour l'analyse automatisée concernent la datation, en particulier sur GB. Selon une étude menée sur un échantillon aléatoire du corpus anglophone de GB, un peu plus de 10% des livres contenaient une erreur de datation, avec presque 38% des livres comprenant une ou plusieurs erreurs de métadonnées (James et Weiss 2012). Geoff Nunberg (2009) fait ainsi semblant de se réjouir de découvrir que 1899, « *annus mirabilis* », a vu entre autres la publication de *Un tueur sous la pluie* de Raymond Chandler ou de *La Condition humaine* d'André Malraux, respectivement parus en 1935 et 1933.

[47] On retrouve par exemple sur Gallica des erreurs de transcription, en particuler sur les césures en fin de ligne avec des formes comme « ab.-■ :sorber;; » ou « dôve- ' loppement ». Sauf à chercher un terme entre guillemets, l'outil de recherche de Gallica inclut d'ailleurs de légères variations dans ses résultats de manière à prendre en compte ce problème de faux négatifs.

[48] Quelques plateformes comme Eureka.cc (n.d.) proposent d'utiliser des opérateurs booléens avancés, par exemple avec l'opérateur « % » accompagné d'un chiffre pour ajouter un critère de proximité entre deux mots-clés (e.g. « %4 » pour quatre mots maximum de séparation). La plupart des plateformes ne fournissent toutefois que quelques opérateurs booléens de base, souvent mal indiqués, comme dans le cas de Gallica (c2014) qui ne propose à l'utilisation que les opérateurs « ET », « OU », « SAUF » et « " " ».

partir des données de GNV. Pour n'en retenir qu'un seul cas, la baisse constatée de la fréquence d'utilisation du mot « *men* » dans le corpus anglophone de GNV ne peut par exemple en aucun cas être simplement expliquée à l'aune d'une « bataille des sexes », en comparaison de l'évolution positive de l'utilisation du mot « *women* » (nous traduisons). En effet, cette hypothèse, qui suppose que davantage d'attention serait désormais portée aux femmes et une moindre attention aux hommes ne prend pas en compte plusieurs facteurs comme a) le fait que parler davantage des femmes ne suppose pas d'en parler de manière valorisante et autonomisante (*empowering*), comme en témoigne les efforts de réessentialisation de la différence de genre qui guident une partie conséquence de la littérature médicale occidentale du XIX^{e} siècle (Fraisse [1989] 1995 ; Laqueur [1990] 1992), b) le fait est que l'utilisation à portée faussement généralisante du mot « *men* » a été critiquée par le mouvement féministe au moins à partir des années 1970 en ce qu'elle invisibilise et exclut les femmes de la sphère publique et donc qu'une baisse de fréquence peut également signifier également une utilisation croissante de termes plus inclusifs comme « *humans*[49] » en lieu et place de « *men* » (Spender [1980] 1985 ; Warren [1986]), et enfin c) le fait qu'une diminution en proportion ne signifie pas une diminution en nombre absolu et qu'il est donc possible qu'une proportion plus faible *et* qu'un nombre plus important de pages soient dans le même temps consacrés aux hommes, puisque le nombre total de mots du corpus anglophone augmente de manière relativement continue entre 1800 et 2000 (Pechenick, Danforth et Dodds 2015, 2). En bref, la comparaison des deux courbes ne permet strictement aucune conclusion à elle seule et ne peut prétendre constituer « un nouveau type de preuve en sciences humaines » (Michel et al. 2011, 181), puisqu'elle dépend toujours d'une interprétation qui lui est indépendante. Ainsi que le résument Eitan A. Pechenick, Christopher M. Danforth et Peter S. Dodds (2015, 23),

[49] À comparer sur GNV entre 1950 et 2000 (période où l'évolution est la plus marquée) la fréquence d'utilisation des mots « *man* », « *human* » et « *person* » (au singulier et au pluriel, avec et sans majuscule, en incluant des pondérations pour faciliter la comparaison des courbes) avec la requête « ((humans+Humans)*20), ((human+Human)*2), (men+Men), (man+Man), ((person+Person)*2), ((persons+Persons)*3) », on peut constater une augmentation en proportion de l'utilisation de « *humans* », avec une fréquence multipliée par sept ou huit en un demi-siècle, une très légère augmentation de la fréquence d'utilisation de « *human* », et une baisse d'environ 40% de la fréquence d'utilisation de « *man* » et « *men* ». Ce qui pourrait aller dans le sens de notre hypothèse mais nécessiterait d'être confirmé par d'autres moyens. L'évolution des usages de « *person* » et « *persons* » montre d'ailleurs les limites de ces conclusions puisqu'elle suit une tendance inverse, avec une légère baisse au pluriel là où on pourrait attendre l'inverse.

GB constitue donc « au mieux un *proxy* limité » pour accéder à des marqueurs différés d'évolutions sociétales et nécessite « une approche très précautionneuse » dans toute utilisation[50].

On peut remarquer que les prétentions des chercheurs de Google reposent en bonne partie sur l'idée que la recherche proposée serait impossible à répliquer de manière non-automatisée du fait du temps nécessaire à la lecture de milliards de mots. À la comparaison absconse déjà évoquée entre les données de Walmart et celles de la bibliothèque du Congrès(cf. 1.3) s'ajoute donc l'idée que la lecture intégrale d'un corpus non-représentatif serait pour une raison obscure nécessaire à la production d'une recherche « objective », affirmation répétée d'article en article : il faudrait 80 ans à lire 200 mots par minute pour lire la production anglophone du corpus de GNV pour la seule année 2000, insistent par exemple Michel et al. (2011, 176), entre autres comparaisons impliquant la longueur du génome humain ou la distance de la Terre à la Lune. Pourtant, toute personne ayant déjà utilisé un moteur de recherche saura qu'elle n'a pas eu à lire l'intégralité du contenu indexé pour obtenir réponse à sa requête, l'outil visant précisément à éviter cela. On remarquera enfin que parler de « lecture » pour une analyse automatisée renvoie une fois de plus à une anthropomorphisation des outils informatiques et participe d'une déconsidération pour l'éventualité même de l'analyse qualitative d'un ensemble massif de données.

Dans les articles consacrés au sujet, la comparaison avec l'approche qualitative, lorsqu'elle n'est pas simplement l'éléphant dans la pièce, se trouve en effet la plupart du temps prestement écartée par la mobilisation d'une distinction notionnelle introduite au début des années 2000 par Franco Moretti entre la « lecture distante » (*distant reading*) et la « lecture attentive » (*close reading*). Le théoricien littéraire italien prophétise alors l'avènement de la « lecture distante », lecture « "de seconde main" » originellement introduite comme « un patchwork des lectures de recherches [produites par] d'autres personnes, *sans une seule lecture*

[50] Ces approches existent en HN, par exemple dans les travaux de Ted Underwood, qui utilise au mieux les méthodes des HN en tentant par exemple avec David Bamman d'identifier des indicateurs de l'évolution des caractérisations genrées dans la littérature anglophone (Underwood et Bamman 2016). Notre analyse des risques associés à un emploi sans précaution des méthodes des HN ne signifie donc pas leur rejet, mais plutôt un rappel de l'absence de solution technologique miracle et de la pertinence d'approches complémentaires, y compris qualitatives.

textuelle directe ». L'idée centrale de la « lecture distante » est alors le postulat selon lequel « plus le projet est ambitieux, plus la distance au texte doit être importante » (Moretti [2000] 2013, 48, nous traduisons). De fait, par sa recherche d'un biais méta-analytique d'appréhension de phénomènes littéraires, la méthode de Moretti présente des affinités fortes avec les contraintes techniques d'une analyse automatisée. Elle vise ainsi à produire des représentations graphiques par une synthèse exhaustive de la production littéraire étudiée, privilégie la corrélation à la causalité (Moretti [2006] 2013, 142-144), présente « une préférence claire pour l'explication des structures générales par rapport à l'interprétation de textes individuels » (Moretti 2005, 91, nous traduisons), dont elle ignore le contenu de la même manière que l'analyse automatisée se limite à la construction de nouvelles métadonnées et/ou à l'analyse des métadonnées existantes (Pustejovsky et Stubbs 2012). Certaines critiques adressées à l'analyse automatisée sont d'ailleurs transposables presque telles quelles à l'approche de Moretti malgré l'absence de déterminant informatique[51], en particulier sa méthode supposée être la seule à même de permettre d'analyser exhaustivement un corpus et supposant cette tâche nécessaire (Moretti [2000] 2013, 45-46), ainsi que l'illusion d'objectivité entretenue par la systématisation de la tâche de synthèse et renforcée par un résultat graphique qui en invisibilise les étapes de production[52]. Le parallèle est d'autant plus aisé à formuler que Moretti ([2008] 2013, 164; [2011] 2013, 212) trouve lui-même dans le développement de bases

[51] Plutôt que de « fétichisme de la technologie », Josiane Jouët (2011, 80) parle d'ailleurs plus largement de « "quantophrénie" techniciste », en référence à la critique de Pitirim Sorokin ([1956] 2008, 108) à l'encontre de l'application de méthodes quantitatives à des phénomènes qui ne s'y prêtent pas.

[52] La représentation graphique constitue certes un outil de synthèse important mais elle n'a pas en elle-même de valeur démonstrative, ce que Moretti (2005, 31) oublie lorsqu'il décide de privilégier la cohérence esthétique en excluant d'une représentation des bornes chronologiques des genres romanesques britanniques des XVIII^e^ et XIX^e^ siècle les romans de détective et de science-fiction. Cette décision, justifiée au nom de leur « durée [de vie] particulièrement longue » qui « semble requérir une approche différente » des autres genres littéraires, tend à montrer que le théoricien littéraire cède à une illusion d'objectivité par une recherche de cohérence graphique, Underwood (2016) ayant montré par ailleurs qu'il pouvait être cohérent de les considérer comme des genres à part entière. L'approche de Moretti (2005, 30) s'avère de fait ambiguë quant à sa position dans la tension entre l'idée que « les données quantitatives sont utiles puisqu'elles sont indépendantes de l'interprétation » et le constat inverse qu'elles « exigent souvent une interprétation qui transcende le domaine quantitatif ». D'une part, le théoricien littéraire a en effet bien conscience que ses données ne sont en rien objectives, citant notamment la nécessité de trancher certains conflits de datation ou la question de la place différenciée attribuée à la production romanesque féminine dans les études sur lesquelles il se base (Moretti 2005, 18, 26-27). Mais, il cherche tout de même à en tirer une conjecture sur l'évolution des genres littéraires sur un modèle pseudo kuhnien et estime donc à vingt-cinq ou trente ans la durée de vie d'un genre littéraire (Moretti 2005, 20). Or, même à écarter le contre-exemple des romans de détective et de science-fiction, son approche ne permet que de mesurer l'évolution des *assignations* de romans à des genres spécifiques

de données numérisées et dans l'introduction de GNV une confirmation de l'intérêt de sa méthode et que ses textes sur le sujet constituent aujourd'hui l'une des principales références en HN[53].

Quant à l'invisibilisation d'une alternative qualitative à l'analyse des grands ensembles de données, enfin, l'opposition fondamentale qui guide le théoricien littéraire ne porte pas sur les lectures attentive et distante, mais plutôt sur les analyses qualitative et quantitative des contenus. Moretti (2013, 44) le reconnait d'ailleurs puisque, revenant sur les conditions dans lesquelles il a commencé à parler de « lecture distante », il mentionne avoir d'abord pensé parler de de « lecture sérielle » (*serial reading*), par « allusion aux procédures de base de l'histoire quantitative ». Ce qui se voulait en premier lieu une « blague » a pourtant eu pour conséquence de dichotomiser artificiellement les approches, par l'invisibilisation d'une réponse complémentaire à un même problème[54]. En effet, à opposer une « lecture distante » à une « lecture attentive » considérée « dans toutes ses incarnations, de la *new criticism* à la déconstruction », Moretti ([2000] 2013, 48) se comporte comme si l'ensemble de l'analyse de texte préexistante relevait de la lecture attentive. Or, on peut notamment opposer dans ses objectifs, avec Michael Leff (1992, 223), une « critique textuelle », à laquelle se raccroche la lecture attentive, et une « critique idéologique », qui ouvre à une solution alternative (nous traduisons). Le rapport au texte sera en effet très différent selon qu'on considère celui-ci comme présentant des « caractéristiques "intrinsèques" » (Jasinski 2001, 91, nous traduisons)

puisqu'elle constitue une méta-analyse des études publiées sur le sujet. Détourner ce résultat pour parler des genres littéraires revient alors à faussement objectiver ce qui ne constitue jamais qu'un ensemble de conventions déterminées à des fins de classification, de publicité et/ou d'analyse, comme le montre par contraste la classification surspécialisée utilisée sur plateforme Netflix (cf. Amatriain 2013). Enfin, Moretti s'abstient d'utiliser des outils de représentation du type de la boîte à moustaches, et ses bornes chronologiques ne peuvent donc être que très imprécises, un roman précurseur ou tardif allongeant artificiellement la durée de vie mesurée.

[53] Moretti se distancie toutefois des prétentions hégémoniques des HN en affirmant la complémentarité des approches de lecture « attentive » et « distante » et en refusant de voir dans les HN la solution à ce qu'il considère à tort comme une impasse des SHS, à savoir l'incapacité à produire des théories comparables dans leur forme à celles des SNF (Dinsman et Moretti 2016).

[54] Plus exactement, Moretti (2005, 7) mentionne ce qu'il nomme une « lecture "extensive" » par opposition à une lecture intensive, mais ne semble pas considérer qu'une approche rigoureuse lui soit associable, évoquant le fait de « lire de nombreux textes une fois et superficiellement » (nous traduisons).

ou comme dépendant d'une « formation idéologique plus large », nécessaire à prendre en compte pour que « le "texte" entier émerge » (Leff 1992, 224-225).

C'est donc la conception de ce que représente un texte qui diffère (Leff 1992, 225), au même titre que la « lecture distante » de Moretti ([2000] 2013, 53) se distingue à considérer le livre comme l'unité indivisible soumise à analyse, c'est-à-dire comme une boîte noire dont on ne connait pas le fonctionnement mais dont on connait les caractéristiques. De « lecture » il n'est en effet pas question dans la « lecture distante », qui implique de ne recourir qu'à des sources secondaires, en particulier des « études critiques indépendantes », en « appren[ant] à *ne pas* lire » pour concentrer son attention sur des ensembles plus larges comme le genre littéraire, à la recherche de motifs récurrents (Moretti [2000] 2013, 48-49, 52).

3.4 Pour une lecture intertextuelle des données massives

Ce critère de distinction par l'appréhension de l'objet-texte nous semble permettre d'expliquer pourquoi une lecture qualitative d'ensembles massifs de données est tout à fait possible et même souhaitable. On essaiera de le montrer à présent à travers les analyses de Julia Kristeva, Roland Barthes et Michel Foucault, et la caractérisation de ce qu'on appellera par praticité la « lecture intertextuelle », par contraste avec la lecture attentive et la lecture distante. Celle-ci a pour différence majeure avec ces deux approches de considérer le texte comme un objet qui n'existe comme tel qu'en relation avec d'autres textes, de la même manière qu'un individu n'existe comme individu que par sa socialisation à d'autres individus. « [T]out texte se construit comme mosaïque de citations, tout texte est absorption et transformation d'un autre texte », explique ainsi Kristeva ([1966] 1969, 87). Si l'écriture devient « absorption et réplique » à des textes préexistants, alors s'installe une « ambivalence », où l'intentionnalité auctoriale devient secondaire (Kristeva [1966] 1969, 89), au point où Barthes ([1968] 1984) pense – un peu vite – pouvoir en dresser l'acte de décès. Cette approche s'oppose de facto à une lecture attentive en ce que, plutôt que d'être considéré dans une unité intrinsèque, le texte est étudié « comme une intertextualité », « pens[é] ainsi dans (le texte de) la société et l'histoire » (Kristeva 1968, 104). Le « texte général » dont il est fait étude est alors d'une nature

sociale et politique, qu'il s'agisse de « la culture » pour Kristeva (1968, 103) ou du « régime de vérité » pour Foucault ([1977] 2001a, 158-159).

Concrètement, il s'agit « à la fois de distinguer les événements[55], de différencier les réseaux et les niveaux auxquels ils appartiennent, et de reconstituer les fils qui les relient et les font s'engendrer les uns à partir des autres », explique Foucault ([1977] 2001a, 144-145). La « différence » d'un texte, sa spécificité, « s'articule sur l'infini des textes, des langages, des systèmes », et il s'agira de « remettre chaque texte, non dans son individualité, mais dans son jeu, le faire recueillir, avant même d'en parler, par le paradigme infini de la différence, le soumettre d'emblée à une typologie fondatrice, à une évaluation » (Barthes 1970, 11). Cette lecture converge avec la « lecture distante » sur le refus du respect d'un corpus canonique, à travers en particulier l'intérêt de Foucault (1997, 8-10) pour les « savoirs assujettis », qui représentent soit des « savoirs ensevelis de l'érudition », « masqués à l'intérieur des ensembles fonctionnels et systématiques », soit des « savoirs disqualifiés par la hiérarchie des connaissances et des sciences », considérés « non conceptuels, […] insuffisamment élaborés ». Le « couplage » de ces deux sources de connaissance, c'est-à-dire l'abandon de tout canon, constitue même le fondement de son « projet généalogique ».

Pour en rester à l'exemple foucaldien, ce rejet des canons implique, pour être mis en pratique, de remplacer le principe d'exhaustivité de la lecture distante par un principe de saturation[56].

[55] Le terme « événement » est ici utilisé dans un sens assez large proche de celui que Barthes donne au terme « texte », Foucault ([1978] 2001, 465) expliquant par ailleurs qu'« il faut considérer le discours comme une série d'événements, comme des événements politiques, à travers lesquels du pouvoir est véhiculé et orienté ».

[56] Cette idée de saturation n'est pas assumée ouvertement par Foucault, malgré une prise en compte croissante de la subjectivité de ses choix documentaires à partir du début des années 1970 (Revel 2002, 16). Elle est donc le fait d'une extrapolation inspirée de la théorisation ancrée (*grounded theory*), mais nous semble cohérente avec l'approche du théoricien français. Certes, Foucault ([1966] 2001, 527) affiche une aspiration exagérée à l'exhaustivité lorsqu'il explique qu'« [i]l faut pouvoir tout lire, connaître toutes les institutions et toutes les pratiques » pour pouvoir tout vérifier soi-même, mais il reconnait alors à demi-mots la nécessité de choix « inavouables, et [qui] ne doivent pas exister » et mentionne plus tard la dimension subjective du choix de certaines archives (Foucault [1977] 2001b, 237). De plus, Foucault ([1975] 2001, 1609) note également que, dans l'étude d'un thème large, « le corpus est en un sens indéfini : on n'arrivera jamais à constituer l'ensemble des discours sur la folie, même en se limitant à une époque donnée, dans un pays donné ». Une manière de réconcilier ces trois affirmations est donc de remplacer, dans la méthode affichée, l'aspiration foucaldienne à l'exhaustivité par un principe de saturation plus proche de la pratique effective. Celui-ci suppose, en cohérence avec les principes de la lecture intertextuelle, qu'existe un « point de saturation, qu'il faut bien en-

La prétention n'est dès lors plus celle d'une exhaustivité dans la lecture mais bien d'une exhaustivité dans les catégories de contenus pertinentes à la problématique de recherche. Plutôt qu'à une représentativité statistique, c'est donc une représentativité thématique qui est visée, en « valorisant la variation par rapport à la quantité » et en laissant de côté toute évaluation de fréquence (Morse 1995, 147, nous traduisons). La recherche qualitative sur des DM n'est dans ce cadre aucunement problématique, puisque « la logique de saturation des différences n'a aucun rapport géométrique avec la population globale (le corpus potentiel) » (Atifi et al. 2006, 168). Dès lors, il n'est plus nécessaire de « lire » l'intégralité des données disponibles, mais simplement de disposer d'une méthode à même d'y sélectionner de manière rigoureuse l'ensemble des contenus pertinents. Au lieu d'une complexification par la surabondance de données, le cadre numérique offre donc au contraire une simplification de l'accès aux données, en particulier via la recherche par mots-clés sur des documents océrisés disponibles en ligne. Les problèmes d'infrastructure qui empêchent une véritable exploitation quantitative des DM en SHS sont ici secondaires puisque l'océrisation, même imparfaite, démultiplie tout de même les capacités de recherche sans impact sur le seuil de saturation. De même, l'identification de documents non-numérisés pertinents à l'analyse, importante pour compléter le corpus faussement complet des DM, se trouve grandement facilitée, à la fois a) via les documents eux-mêmes, puisque c'est l'objectif même de la lecture intertextuelle que de les mettre au jour, et b) via les différentes bases de données, qui tendent à être adossées ou fusionnées à un ou plusieurs catalogues, comme HathiTrust Digital Library avec Worldcat, ou Gallica avec Data BNF.

Conclusion

La spécificité de l'approche qualitative la plus pertinente du point de vue de la défense d'une lecture intertextuelle, puisqu'irréplicable de manière automatisée, s'avère enfin la capacité à produire une interprétation sémantique. Dans l'approche intertextuelle, en effet, le texte se trouve caractérisé précisément par le maintien d'une stricte distinction entre le contenu et le

tendu largement dépasser pour être assuré de la validité de ses conclusions » (Bertaux 1980, 206), au-delà duquel aucune nouvelle donnée ne permet de tirer de conclusion nouvelle significative (Glaser et Strauss 1967, 61).

support qui signifie l'impossibilité à le saisir de manière procédurale et computationnelle : « Une œuvre [i.e. un support] est un objet fini, computable, qui peut occuper un espace physique », là où « le texte est un champ méthodologique », résume ainsi Barthes ([1973]). À la lecture, « il n'y a jamais un *tout* du texte […] : il faut à la fois dégager le texte de son extérieur et de sa totalité » (Barthes 1970, 13). En d'autres termes, la lecture devient une tâche aussi importante – et vertigineuse – que l'écriture, avec pour ambition irréalisable de situer un texte à la fois dans et hors de son support. Pour cette raison, elle demeure « toujours inachevée » et l'interprétation ne mène jamais à une vérité de « la chose qui s'offre à l'interprétation », tout simplement parce qu'« [i]l n'y a rien d'absolument premier à interpréter » (Foucault [1967] 2001, 598-599). « Interpréter un texte, ce n'est pas lui donner un sens […], c'est au contraire apprécier de quel pluriel il est fait », explique Barthes (1970, 13). De plus, dans une seconde composante essentielle, « l'interprétation se trouve devant l'obligation de s'interpréter soi-même à l'infini; de se reprendre toujours » (Foucault [1967] 2001, 601). L'herméneutique implique donc un rapport éminemment subjectif au texte, puisque l'interprétation sera toujours située en fonction de la personne qui la formule et ne pourra jamais prétendre atteindre une vérité absolue, puisqu'elle acte le caractère chimérique de la recherche de celle-ci[57].

De manière plus directement applicable à l'analyse, Paul Ricœur (1986, 151-160) parle d'un « *arc herméneutique* » pour désigner la continuité entre d'une part le fait de « se mettre dans le sens indiqué par [une] relation d'interprétation portée par le texte lui-même », c'est-à-dire « une interprétation objective, et en quelque sorte intratextuelle », et d'autre part « l'interprétation comme appropriation », c'est-à-dire à la fois comme « victoire sur la distance culturelle » au texte et comme « fusion de l'interprétation du texte à l'interprétation de soi-même ». Même si c'est dans une certaine mesure le cas chez Barthes, il ne s'agit donc pas de laisser s'exprimer un arbitraire, mais de simplement reconnaître la limite à une explication à

[57] Cette conception de la subjectivité ne suppose pas d'ignorer les déterminants sociaux, même si ces derniers demeurent souvent au second plan de l'analyse de Foucault ou de Barthes. Susan Stanford Friedman (1991) propose d'ailleurs, en réponse à ce manque, une réécriture de la théorie foucaldienne intéressée plus directement par la prise en compte des rapports de domination. Cette approche modifie toutefois le « texte général », puisque ce dernier sera dès lors ce qu'on peut appeler un « sous-régime de vérité », variante du régime de vérité qui pourra éventuellement entrer en contradiction partielle avec d'autres élément de l'ensemble.

prétention objective. Or, ce biais indépassable de l'interprétation n'a pas de raison de n'être pas transmis à des créations algorithmiques, y compris fondées sur l'apprentissage machinique[58].

Au contraire d'une nécessité, le traitement automatisé des contenus sémantiques apparait ainsi difficile à appréhender, dès lors qu'on considère les données et les simulations pour ce qu'elles sont et non plus pour des reproductions fidèles de la réalité. Considérer le texte comme « pris dans un système de renvois à d'autres livres, d'autres textes, d'autres phrases », c'est-à-dire comme « nœud dans un réseau » (Foucault 1969, 34), n'implique en effet pas forcément que les ressources théoriques, matérielles et logicielles nécessaires à un traitement algorithmique puissent être réunies. Comme le résume Aurélien Bénel (2014), « l'herméneutique se distingue des mirages du positivisme par la prise en compte de *documents* plutôt que de "données", d'*interprétations* plutôt que d'inférences et de *débats intersubjectifs* plutôt que de "preuves" », et le terrain numérique actuel s'avère à ce titre particulièrement adapté à l'approche qualitative.

[58] Aylin Caliskan, Joanna J. Bryson et Arvind Narayanan (2017, 185) estiment par exemple que « si nous construisons un système intelligent qui en apprend assez des propriétés du langage pour être capable de le comprendre et de le produire, il acquerra aussi dans le processus des associations culturelles historiques, dont certaines peuvent être questionnables », et l'automatisation d'un processus de sélection de CV ne mettra donc pas fin aux biais racistes ou sexistes constatables aujourd'hui (nous traduisons). Le risque de voir les DM être de facto mises au service de décisions discriminatoires, même de manière non intentionnelle (Barocas et Selbst 2016), a d'ailleurs été confirmé par une enquête de *ProPublica* sur les biais racistes du logiciel Compas utilisé dans plusieurs États étasuniens pour prédire le risque de récidive criminelle (Angwin et al. 2016). Caliskan, Bryson et Narayanan (2017, 185) recommandent en conséquence logiquement d'associer aux algorithmes des « caractérisations explicites des comportements acceptables ».

Bibliographie

Ackoff, Russell L. [1974] 1997. « Systems, Messes and Interactive Planning. » In *The Social Engagement of Social Science: A Tavistock Anthology Volume 3: The Socio-Ecological Perspective*, sous la dir. de Eric L. Trist, Fred E. Emery, Hugh Murray et Beulah Trist, 417-438. Philadelphie: University of Pennsylvania Press. Consulté le 12 décembre 2017. http://www.moderntimesworkplace.com/archives/ericsess/sessvol3/Ackoffp417.opd.pdf.

Amatriain, Xavier. 2013. « Big & Personal: Data and Models behind Netflix Recommendations. » 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Chicago, 11-14 août 2013. Consulté le 24 décembre 2017. https://pdfs.semanticscholar.org/4b89/dfc8782bf0fc403b5b43e5e55ea1d4dc44c6.pdf.

Amiard-Chevrel, Claudine. 1990. « Frères ennemis ou faux frères ? : Théâtre et cinéma avant le parlant. » In *Théâtre et cinéma années vingt : Une quête de modernité*, sous la dir. de Claudine Amiard-Chevrel, 9-32. Lausanne (Suisse): L'Âge d'homme.

Anderson, Chris. 2008. « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. » *Wired*, 23 juin 2008. Consulté le 3 décembre 2016. https://www.wired.com/2008/06/pb-theory/.

Anguelov, Dragomir, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent et Josh Weaver. 2010. « Google Street View: Capturing the World at Street Level. » *Computer* 43 (6). https://static.googleusercontent.com/media/research.google.com/fr//pubs/archive/36899.pdf.

Angwin, Julia, Jeff Larson, Surya Mattu et Lauren Kirchner. 2016. « Machine Bias: There's Software Used Across the Country to Predict Future Criminals and It's Biased Against Blacks. » *ProPublica*, 23 mai 2016. Consulté le 6 octobre 2017. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arnaud, Pascal, Anne Baillot, Aurélien Berra, Dominique Boullier, Thomas Cauvin, Georgios Chatzoudis, Arianna Ciula, Camille Desenclos, André Donk, Marten Düring, Natalia Filatkina, Sascha Foerster, Sebastian Gießmann, Martin Grandjean, Franziska Heimburger, Christian Jacob, Mareike König, Marion Lamé, Lilian Landes, Matthias Lemke, Anika Meier, Benoît Majerus, Claudine Moulin, Pierre

Mounier, Marc Mudrak, Cynthia Pedroja, Jean-Michel Salaün, Markus Schnöpf, Julian Schulz, Bertram Triebel et Milena Žic-Fuchs. 2013. *Jeunes chercheurs et humanités numériques : Un manifeste*. Institut historique allemand. Consulté le 19 novembre 2017. http://dhiha.hypotheses.org/1108.

Atifi, Hassan, Christophe Lejeune, Goritsa Ninova et Manuel Zacklad. 2006. « Méthodologie transdisciplinaire de gestion de corpus pour les disciplines de l'interaction : Recherche de principes directeurs. » Corpus en lettres et sciences sociales : Des documents numériques à l'interprétation, Albi (France), 10-14 juillet 2006. Consulté le 26 décembre 2017. http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Sommaire.html.

Atkinson, John. 2009. « Intelligent Text Mining: Putting Evolutionary Methods and Language Technologies Together. » In *Handbook of Research on Text and Web Mining Technologies*, sous la dir. de Min Song et Yi-Fang Brook Wu, 37-59. Hershey (Pennsylvanie), New York, Londres: Information Science Reference.

Babinet, Gilles. 2015. *Big Data, penser l'homme et le monde autrement*. Paris: Le Passeur.

Barocas, Solon et Andrew D. Selbst. 2016. « Big Data's Disparate Impact. » *California Law Review* 104 (3): 671-732. https://pdfs.semanticscholar.org/1d17/4f0e3c391368d0f3384a144a6c7487f2a143.pdf.

Barthes, Roland. 1964. « Éléments de sémiologie. » *Communication* (4): 91-135. http://www.persee.fr/doc/comm_0588-8018_1964_num_4_1_1029.

———. 1970. *S/Z*. Paris: Le Seuil.

———. [1968] 1984. « La mort de l'auteur. » In *Le Bruissement de la langue : Essais critiques IV*, 54-60. Paris: Le Seuil.

———. [1973]. Théorie du texte. In *Encyclopædia Universalis*. Consulté le 26 septembre 2016. http://asl.univ-montp3.fr/e41slym/Barthes_THEORIE_DU_TEXTE.pdf.

Baudrillard, Jean. 1981. *Simulacres et simulation*. Paris: Galilée.

Becker, Howard S. [1986] 2007. *Writing for Social Scientists: How to Start and Finish Your Thesis, Book, or Article*, 2nde éd. rev. et augm. Chicago, Londres: The University of Chicago Press.

Bénel, Aurélien. 2014. « Quelle interdisciplinarité pour les "humanités numériques" ? » *Les Cahiers du numérique* 10 (4): 103-132. https://www-cairn-info.proxy.bibliotheques.uqam.ca:2443/revue-les-cahiers-du-numerique-2014-4-page-103.htm.

Berry, David M. 2011a. « The Computational Turn: Thinking About the Digital Humanities. » *Culture Machine* (12). http://www.culturemachine.net/index.php/cm/article/view/440/470.

———. 2011b. *The Philosophy of Software: Code and Mediation in the Digital Age*. Basingstoke (Angleterre), New York: Palgrave Macmillan.

Bertaux, Daniel. 1980. « L'approche biographique : Sa validité méthodologique, ses potentialités. » *Cahiers internationaux de sociologie* (69): 197-225. http://www.jstor.org/stable/40689912.

Biemann, Chris, Gregory R. Crane, Christiane D. Fellbaum et Alexander Mehler. 2014. « Computational Humanities: Bridging the Gap Between Computer Science and Digital Humanities. » *Dagstuhl Reports* 4 (7): 80-85. https://www.researchgate.net/publication/318882169_Computational_Humanities_-_bridging_the_gap_between_Computer_Science_and_Digital_Humanities_Dagstuhl_Seminar_14301.

BNF. c2013. *Le web de données à la BNF*. Bibliothèque nationale de France. Consulté le 8 décembre 2017. http://www.bnf.fr/fr/professionnels/web_donnees_applications_bnf/a.web_donnees_bnf_intro.html.

Bouchet, Thomas, Guillaume Carnino et François Jarrige. 2016. « L'université face au déferlement numérique. » *Variations* (19). http://journals.openedition.org/variations/740.

Boudon, Raymond. 1968. « La machine à accélérer le temps : Comment tous les outils, les ordinateurs ne valent que par la manière dont ils sont utilisés. » *Le Nouvel Observateur*, 8 mai 1968, 36-38. Consulté le 18 décembre 2017. http://referentiel.nouvelobs.com/archives_pdf/OBS0182_19680508/OBS0182_19680508_036.pdf ; http://referentiel.nouvelobs.com/archives_pdf/OBS0182_19680508/OBS0182_19680508_037.pdf ; http://referentiel.nouvelobs.com/archives_pdf/OBS0182_19680508/OBS0182_19680508_038.pdf.

Bourdieu, Pierre. 1978. « Sur l'objectivation participante : Réponse à quelques objections. » *Actes de la recherche en sciences sociales* (23): 67-69. http://www.persee.fr/docAsPDF/arss_0335-5322_1978_num_23_1_2609.pdf.

———. 1979. *La Distinction : Critique sociale du jugement*. Paris: Minuit.

Boyadjian, Julien. 2014. « Twitter, un nouveau "baromètre de l'opinion publique" ? » *Participations* 8 (1): 55-74. http://www.cairn.info.proxy.bibliotheques.uqam.ca:2048/revue-participations-2014-1-page-55.htm.

boyd, danah michele et Kate Crawford. [2011] 2012. « Six provocations à propos des *big data*. » In *Read/Write Book 2 : Une introduction aux humanités numériques*, sous la dir. de Pierre Mounier, 197-219. Marseille: OpenEdition Press. Consulté le 8 décembre 2017. http://books.openedition.org/oep/273?lang=fr.

Bricka, Ivan. 2016. « Le débat sur la conceptualité des contenus de la perception : Une critique merleau-pontienne. » département de philosophie, université du Québec à Montréal. Consulté le 27 septembre 2016. http://www.archipel.uqam.ca/8567/1/M14223.pdf.

Burnard, Lou. [2009] 2012. « Du *literary and linguistic computing* aux *digital humanities* : Retour sur 40 ans de relations entre sciences humaines et informatique. » In *Read/Write Book 2 : Une introduction aux humanités numériques*, sous la dir. de Pierre Mounier, 45-58. Marseille: OpenEdition Press. Consulté le 5 décembre 2017. http://books.openedition.org/oep/242.

Bush, Vannevar. 1945. « As We May Think. » *The Atlantic*, juillet 1945. Consulté le 3 décembre 2016. http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

Butler, Declan. 2013. « When Google Got Flu Wrong. » *Nature* (494): 155-156. https://www.nature.com/news/when-google-got-flu-wrong-1.12413.

Caliskan, Aylin, Joanna J. Bryson et Arvind Narayanan. 2017. « Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. » *Science* 356 (6334): 183-186. http://science.sciencemag.org/content/356/6334/183.

Carlier, Denis. 2015. « Généalogie de l'efficacité ressentie : Étude de l'évolution du rapport gestionnaire à l'efficacité dans le métro de Montréal, 1966-2014. » Centre Urbanisation, culture et société, Institut national de la recherche scientifique ;

Université du Québec à Montréal. Consulté le 25 novembre 2015. http://espace.inrs.ca/2687/.

Chartier, Roger. 1988. « Textes, imprimés, lectures. » In *Pour une sociologie de la lecture : Lectures et lecteurs dans la France contemporaine*, sous la dir. de Martine Poulain, 11-28. Paris: Cercle de la Librairie.

Cook, Samantha, Corrie Conrad, Ashley L. Fowlkes et Matthew H. Mohebbi. 2011. « Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. » *PLoS ONE* 6 (8). http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0023610&type=printable.

Cotte, Dominique. 2004. « Le concept de "document numérique". » *Communication et langages* (140): 31-41. http://www.persee.fr/docAsPDF/colan_0336-1500_2004_num_140_1_3265.pdf.

Courtial, Florence et Angeline Lavigne. 2011. « Presse ancienne régionale : Les enjeux de la numérisation. » *Tire-Lignes : La vie du livre en Midi-Pyrénées*, avril 2011, 16. Consulté le 6 décembre 2017. http://fr.calameo.com/read/0004104457c390e70712b.

Crowley, John. 2003. « Usages de la gouvernance et de la gouvernementalité. » *Critique internationale* (21): 52-61. http://www.cairn.info.proxy.bibliotheques.uqam.ca:2048/article.php?ID_ARTICLE=CRII_021_0052.

Dacos, Marin. 2011. *Manifeste des digital humanities*. Centre pour l'édition électronique ouverte. Consulté le 12 octobre 2017. http://tcp.hypotheses.org/318.

Dacos, Marin et Pierre Mounier. 2015. *Humanités numériques : État des lieux et positionnement de la recherche française dans le contexte international*. Paris: Institut français. Consulté le 20 novembre 2017. https://hal.archives-ouvertes.fr/hal-01228945/document.

de Certeau, Michel. [1980] 1990. *L'Invention du quotidien : Arts de faire*, Nouv. éd. Paris: Gallimard.

de Saussure, Ferdinand. [1916] 1995. *Cours de linguistique générale*, éd. critique. Paris: Payot. Consulté le 26 juillet 2016. http://centenaire-linguistique.org/uploads/medias/downloads/le-cours.pdf.

Delamont, Sara et Paul Atkinson. 2001. « Doctoring Uncertainty: Mastering Craft Knowledge. » *Social Studies of Science* 31 (1): 87-101.

Dépelteau, François. [2000] 2003. *La Démarche d'une recherche en sciences humaines : De la question de départ à la communication des résultats*, 2e éd. Québec: Presses de l'université Laval.

DeVan, Ashley. 2016. *The 7 V's of Big Data*. Impact Radius. Consulté le 5 décembre 2017. https://www.impactradius.com/blog/7-vs-big-data/.

Diebold, Francis X. 2012. A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline. Philadelphie: University of Pennsylvania. Consulté le 2 décembre 2017. http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf.

Dignan, Larry. 2017. « Twitter Q3 Takeaway: Data, Enterprise Tools Can Boost Profits, Revenue. » *ZDNet*, 26 octobre 2017. Consulté le 9 décembre 2017. http://www.zdnet.com/article/twitter-q3-takeaway-data-enterprise-tools-can-boost-profits-revenue/.

Dinsman, Melissa et Franco Moretti. 2016. « The Digital in the Humanities: An Interview with Franco Moretti. » *The Los Angeles Review of Books*, 2 mars 2016. Consulté le 26 décembre 2017. https://lareviewofbooks.org/article/the-digital-in-the-humanities-an-interview-with-franco-moretti.

Émerit-Bibié, Laetitia. 2016. Pourquoi je m'intéresse (encore) à Facebook ? In *Mémo numérique(s) : Mes mots et mémos autour du discours numérique*, sous la dir. de Laetitia Émerit-Bibié. Marseille: Centre pour l'édition électronique ouverte. Consulté le 8 décembre 2017. http://memo.hypotheses.org/169.

Eureka.cc. n.d. Les opérateurs booléens avancés. Kirkland (Washington). Consulté le 18 janvier 2018. https://www.cmaisonneuve.qc.ca/wp-content/uploads/2016/10/Les-operateurs-booleens-de-base-Eureka.pdf.

Evans, Leighton et Sian Rees. 2012. « An Interpretation of Digital Humanities. » In *Understanding Digital Humanities*, sous la dir. de David M. Berry, 21-41. Basingstoke (Angleterre), New York: Palgrave Macmillan.

Feather, John. 2013. « Information Society. » In *Encyclopedia of Philosophy and Social Sciences*, sous la dir. de Byron Kaldis, Ann E. Cudd, Margaret Gilbert, Ian Jarvie, Tony Lawson, Philip Pettit, John Searle, Raimo Tuomela et Stephen Turner, 473-476. Los Angeles, Londres, New Delhi, Singapour, Washington: SAGE.

Feldman, Ronen et James Sanger. 2007. *The Text-Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge (Angleterre), New York, Melbourne, Madrid, Le Cap, Singapour, São Paulo: Cambridge University Press.

Floridi, Luciano. 2008. « Data. » In *International Encyclopedia of the Social Sciences: Cohabitation–Ethics in Experimentation*, 2nde éd., sous la dir. de William A. Jr. Darity, Eduardo Bonilla-Silva, Philip R. Costanzo, Patrick L. Mason, Paula McClain, Donald M. Nonini, David Scott, Theresa Singleton, David Dietrich, J. Alan Kendrick et Ley Killeya-Jones, 234-237. Macmillan Reference: Détroit, New York, San Francisco, New Haven (Connecticut), Waterville (Maine), Londres.

———. 2010. *Information: A Very Short Introduction*. Oxford (Angleterre), New York: Oxford University Press.

Foucault, Michel. 1969. *L'Archéologie du savoir*. Paris: Gallimard.

———. 1997. *"Il faut défendre la société" : Cours au Collège de France. 1976*. Paris: Gallimard/Le Seuil.

———. [1966] 2001. « Michel Foucault, "Les Mots et les Choses". » In *Dits et écrits I : 1954-1975*, 2nde éd. en 2 vol., 526-532. Paris: Gallimard.

———. [1967] 2001. « Nietzsche, Freud, Marx. » In *Dits et écrits I : 1954-1975*, 2nde éd. en 2 vol., 592-607. Paris: Gallimard.

———. [1975] 2001. « Entretien sur la prison : Le livre et sa méthode. » In *Dits et écrits I : 1954-1975*, 2nde éd. en 2 vol., 1608-1621. Paris: Gallimard.

———. [1977] 2001a. « Entretien avec Michel Foucault. » In *Dits et écrits II : 1976-1988*, 2nde éd. en 2 vol., 140-160. Paris: Gallimard.

———. [1977] 2001b. « La vie des hommes infâmes. » In *Dits et écrits II : 1976-1988*, 2nde éd. en 2 vol., 237-253. Paris: Gallimard.

———. [1977] 2001c. « Le jeu de Michel Foucault. » In *Dits et écrits II : 1976-1988*, 2nde éd. en 2 vol., 288-329. Paris: Gallimard.

———. [1978] 2001. « Dialogue sur le pouvoir. » In *Dits et écrits II : 1976-1988*, 2nde éd. en 2 vol., 464-477. Paris: Gallimard.

Fraisse, Geneviève. [1989] 1995. *Muse de la Raison : Démocratie et exclusion des femmes en France*. Paris : Gallimard.

Frické, Martin. 2009. « The Knowledge Pyramid: A Critique of the DIKW Hierarchy. » *Journal of Information Science* 35 (2): 131-142. http://journals.sagepub.com/doi/abs/10.1177/0165551508094050.

Furet, François. 1968. « Comment l'informatique bouleverse les sciences humaines. » *Le Nouvel Observateur*, 8 mai 1968, 35. Consulté le 15 décembre 2017. http://referentiel.nouvelobs.com/archives_pdf/OBS0182_19680508/OBS0182_19680508_035.pdf.

Gallica. c2014. Comment faire une recherche ? In *Gallica*. Paris: Bibliothèque nationale de France. Consulté le 18 janvier 2018. http://gallica.bnf.fr/html/und/comment-faire-une-recherche.

GAO. 2005. Aviation Security: Secure Flight Development and Testing Under Way, but Risks Should Be Managed as System Is Further Developed. Washington: Government Accountability Office. Consulté le 11 décembre 2017. http://www.gao.gov/new.items/d05356.pdf.

Genette, Gérard. 1987. *Seuils*. Paris: Le Seuil.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynette Brammer, Mark S. Smolinski et Larry Brilliant. 2009. « Detecting Influenza Epidemics Using Search Engine Query Data. » *Nature* (457): 1012-1014. https://static.googleusercontent.com/media/research.google.com/fr//archive/papers/detecting-influenza-epidemics.pdf.

Gitelman, Lisa et Virginia Jackson. 2013. « Introduction. » In *"Raw Data" Is an Oxymoron*, sous la dir. de Lisa Gitelman, 1-14. Cambridge (Massachusetts), Londres: The MIT Press.

Giuliani, Emmanuelle. 2005. « Presse : *La Croix* est parmi les premiers journaux numérisés par la BNF. La Bibliothèque nationale de France s'engage dans une ambitieuse campagne de mise en ligne de la presse française. » *La Croix*, 17 février 2005. Consulté le 8 décembre 2017. base de données Eureka.cc.

Glaser, Barney G. et Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New Brunswick (New Jersey), Londres: AldineTransaction.

Google. [c2009] c2012. *Ngram Viewer*. Google. Consulté le 8 décembre 2017. http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

———. n.d. *Thank You For Stopping By*. Google. Consulté le 8 octobre 2017. https://www.google.org/flutrends/about/.

Grady, Nancy, dir. [2015] 2017. *NIST Big Data Interoperability Framework: Volume 1, Definitions*, brouillon de la version 2 à paraître. Gaithersburg (Maryland): National Institute of Standards and Technology.Consulté le 2 décembre 2017. https://bigdatawg.nist.gov/home.php.

Greenberg, Jane. 2003. « Metadata and the World Wide Web. » In *Encyclopedia of Library and Information Science*, 2e éd., sous la dir. de Miriam A. Drake, 1876-1888. New York: Marcel Dekker. Consulté le 6 décembre 2017. https://pdfs.semanticscholar.org/0124/166ba96a39b25f103968ec8a45075d0155e0.pdf.

Hall, Stuart. [1980] 1997. « Codage/Décodage. » *Sociologie de la communication* 1 (1): 59-71. http://www.persee.fr/docAsPDF/reso_004357302_1997_mon_1_1_3832.pdf.

Haraway, Donna. [1988] 2007. « Savoirs situés : La question de la science dans le féminisme et le privilège de la perspective partielle. » In *Manifeste cyborg et autres essais : Sciences, fictions, féminismes*, 107-142. Paris: Exils.

Harding, Sandra. 1992. « Rethinking Standpoint Epistemology: What is "Strong Objectivity"? » *The Centenial Review* 36 (3): 437-470. http://www.jstor.org/stable/23739232.

Harvey, David. 2003. « The Fetish of Technology: Causes and Consequences. » *Macalester International* 13 (7). http://digitalcommons.macalester.edu/cgi/viewcontent.cgi?article=1411&context=macintl.

Heyer, Gerhard et Marco Büchler. 2010. Some Challenges Posed to Computer Science by the eHumanities. In *Informatik 2010 Service Science: Neue Perspektiven für die Informatik*, sous la dir. de Klaus-Peter Fähnrich et Bogdan Franczyk. Leipzig (Allemagne): Gesellschaft für Informatik. Consulté le 2 décembre 2016. http://cs.emis.de/LNI/Proceedings/Proceedings176/524.pdf.

House of Lords. 2007. The EU/US Passenger Name Record (PNR) Agreement Londres: House of Lords. Consulté le 12 décembre 2017. https://pdfs.semanticscholar.org/69f7/cfe0c24713315d72ccbf4560e574ba77f046.pdf.

James, Ryan et Andrew Weiss. 2012. « An Assessment of Google Books' Metadata. » *Journal of Library Metadata* 12 (1): 15-22. http://www.tandfonline.com/doi/abs/10.1080/19386389.2012.652566.

Jasinski, James. 2001. *Sourcebook on Rethorical Key Concepts in Contemporary Rhetorical Studies*. Thousand Oaks (Californie), Londres, New Delhi: Sage.

Jouët, Josiane. 2011. « Des usages de la télématique aux *Internet Studies*. » In *Communiquer à l'ère numérique : Regards croisés sur la sociologie des usages*, sous la dir. de Julie Denouël et Fabien Granjon, 45-90. Paris: Presses des Mines. Consulté le 29 décembre 2017. https://ecole-ident-num.sciencesconf.org/conference/ecole-ident-num/pages/Jouet_2011.pdf.

Katz, Elihu. 1957. « The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis. » *The Public Opinion Quarterly* 21 (1): 61-78. https://www.jstor.org/stable/2746790.

Kettering, Jon R., Richard A. Berk, Lawrence D. Brown, Nicholas P. Jewell, James D. Kuelbs, John Lehoczky, Daryl Pregibon, Fritz Scheuren, J. Laurie Snell, Elizabeth Thompson et Jack Alexander, dir. 1997. *Massive Data Sets: Proceedings of a Workshop*. Washington: National Academy Press.

Khelifi, Hadria. 2017. La lexicométrie : Un outil efficient pour l'analyse du discours. In *Carnet des jeunes chercheurs du Crem*. Marseille: Centre pour l'édition électronique ouverte. Consulté le 15 décembre 2017. http://ajccrem.hypotheses.org/370.

Kitchin, Rob. 2014a. « Big Data, New Epistemologies and Paradigm Shifts. » *Big Data & Society* 1 (1). doi: 10.1177/2053951714528481.

———. 2014b. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Los Angeles, Londres, New Dehli, Singapour, Washington: SAGE.

Kitchin, Rob et Gavin McArdle. 2016. « What makes Big Data, Big Data?: Exploring the Ontological Characteristics of 26 Datasets. » *Big Data & Society* 1-10. http://journals.sagepub.com/doi/pdf/10.1177/2053951716631130.

Kopf, Johannes, Billy Chen, Richard Szeliski et Michael Cohen. 2010. « Street Slide: Browsing Street Level Imagery. » SIGGRAPH2010: The 37th International Conference and Exhibition on Computer Graphics and Interactive Techniques, Los Angeles Convention Center, Los Angeles, 25-29 juillet 2010. Consulté le 14 décembre 2017. https://pdfs.semanticscholar.org/076b/e4b491c2051d848fd3c7fd4504d8255bffe5.pdf.

Koplenig, Alexander. 2015. « The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets: Reconstructing the Composition of the German Corpus in Times of WWII. » *Digital Scholarship in the Humanities* 32 (1): 169-188. http://doi.org/10.1093/llc/fqv037.

Kovač, Miha. 2008. « "Never mind the Web: Here comes the book." Continuity and discontinuity in the fate of reading. » *Logos* 19 (3): 151-158. http://booksandjournals.brillonline.com/content/journals/10.2959/logo.2008.19.3.151.

Kristeva, Julia. 1968. « Le texte clos. » *Langages* (12): 103-125. http://www.jstor.org/stable/41680693.

———. [1966] 1969. « Le mot, le dialogue et le roman. » In *Σημειωτιχή [Sèméiotikè] : Recherches pour une sémanalyse* 85-111. Paris: Le Seuil.

Kuhn, Thomas Samuel. [1962] 1996. *The Structure of Scientific Revolutions*, 3e éd. Chicago, Londres: The University of Chicago Press.

Lalonde, Catherine. 2017. « Trous de mémoire aux archives de BANQ. » *Le Devoir*, 6 juillet 2017. Consulté le 8 décembre 2017. http://www.ledevoir.com/culture/actualites-culturelles/502775/banq-ralentissement-numerique#.

Landow, George P. 1992. *Hypertext: The Convergence of Contemporary Critical Theory and Technology*. Baltimore (Maryland), Londres: John Hopkins University Press.

Laney, Doug. 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. In *Meta Group Research Note*. Stamford (Connecticut): Meta Group. Consulté le 5 décembre 2017. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

———. 2012. *Deja VVVu: Others Claiming Gartner's Construct for Big Data*. Gartner. Consulté le 5 décembre 2017. https://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/.

Langlais, Pierre-Carl. 2017. Les bibliothèques numériques sont-elles représentatives ? In *Sciences communes*, sous la dir. de Pierre-Carl Langlais. Marseille: Centre pour l'édition électronique ouverte. Consulté le 7 décembre 2017. http://scoms.hypotheses.org/799.

Laqueur, Thomas. [1990] 1992. *La Fabrique du sexe*. Paris: Gallimard.

Latour, Bruno. 2008. « La connaissance est-elle un mode d'existence ? : Rencontre au Muséum de James, Fleck et Whitehead avec des fossiles de chevaux. » In *Vie et expérimentation : Peirce, James, Dewey*, sous la dir. de Didier Debaise, 17-44. Paris: Vrin. Consulté le 12 décembre 2017. http://www.bruno-latour.fr/sites/default/files/downloads/99-HANDBOOK-COURT-FR_0.pdf.

Lazer, David et Ryan Kennedy. 2015. « What We Can Learn From the Epic Failure of Google Flu Trends. » *Wired*, 1er octobre 2015. Consulté le 8 octobre 2017. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/.

Lazer, David, Ryan Kennedy, Gary King et Alessandro Vespignani. 2014. « The Parable of Google Flu: Traps in Big Data Analysis. » *Science* 343 (6176): 1203-1205. http://science.sciencemag.org/content/343/6176/1203.

Le Cun, Yann. 2016. L'apprentissage profond : Une révolution en intelligence artificielle. In *Leçon inaugurale*. Collège de France, Paris: Collège de France. Consulté le 25 novembre 2017. https://www.youtube.com/watch?v=OzZoPVjv8iE.

Leff, Michael. 1992. « Things Made by Words: Reflections on Textual Criticism. » *Quaterly Journal of Speech* 78 (2): 223-231. http://www.tandfonline.com/doi/abs/10.1080/00335639209383991.

Lejeune, Christophe. 2017. « Analyser les contenus, les discours ou les vécus ? À chaque méthode ses logiciels ! » In *Les Méthodes qualitatives en psychologie et sciences humaines de la santé*, sous la dir. de Marie Santiago-Delefosse et Maria del Rio Carral, 203-224. Malakoff (France): Dunod.

Lin, Rong-Gong Jr. 2017. « Revenge of Y2K?: A Software Bug Might Have Caused False Alert for Big (and Very Old) Earthquake. » *Los Angeles Times*, 22 juin 2017. Consulté le 14 décembre 2017. http://www.latimes.com/local/lanow/la-me-earthquakesa-earthquake-68-quake-strikes-near-isla-vista-calif-jyhw-htmlstory.html.

Mandelbrot, Benoît B. [1975] 1995. *Les Objets fractals : Forme, hasard et dimension*, 4e éd. revue. Paris: Flammarion.

Marx, Karl. [1867] 1993. *Le Capital : Critique de l'économie politique. Livre premier : Le procès de production du capital*, trad. de la 4e éd. all. Paris: Presses universitaires de France.

Mayer-Schönberger, Viktor et Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, New York: Eamon Dolan, Houghton Mifflin Harcourt.

McLuhan, Marshall. [1964] 2013. *Understanding Media*. Berkeley (Californie): Gingko Press.

McNulty-Holmes, Eileen. 2014. « Understanding Big Data: The Seven V's. » *Dataconomy*, 22 mai 2014. Consulté le 5 décembre 2017. http://dataconomy.com/2014/05/seven-vs-big-data/.

Meeks, Elijah. 2014. An Introduction to Digital Humanities - Bay Area DH. San Bruno (Californie): Youtube. Consulté le 19 novembre 2017. https://www.youtube.com/watch?v=AvZToQSX244.

Merleau-Ponty, Maurice. [1945] 2008. Phénoménologie de la perception. Chicoutimi (Québec): Université du Québec à Chicoutimi. http://classiques.uqac.ca/classiques/merleau_ponty_maurice/phonomenologie_de_la_perception/phonomenologie_de_la_perception.pdf.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Martin A. Nowack et Erez Lieberman Aiden. 2011. « Quantitative Analysis of Culture Using Millions of Digitized Books. » *Science* 331 (6014): 176-182. http://science.sciencemag.org/content/331/6014/176.

Miles, Matthew B. et A. Michael Huberman. [1983] 1994. *Qualitative Data Analysis: An Expanded Sourcebook*, 2nde éd. Thousand Oaks (Californie), Londres, New Dehli: SAGE.

Miller, Jeff. [c2008] 2017. *Earliest Uses of Symbols of Set Theory and Logic*. Lycos. Consulté le 14 décembre 2017. http://jeff560.tripod.com/set.html.

Mordvintsev, Alexander, Christopher Olah et Mike Tyka. 2015. Inceptionism: Going Deeper into Neural Networks. In *Google Research Blog*. Moutain View (Californie): Google. Consulté le 15 décembre 2017. https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

Morelli, Pierre. 2005. « Paratexte et transcendance textuelle dans les supports numériques "offline" : Essai de typologie. » Enjeux et usages des TIC : Aspects sociaux et culturels, Bordeaux, 22-23 septembre 2005. Consulté le 11 décembre 2017. https://hal.archives-ouvertes.fr/halshs-00111708/document.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Londres, New York: Verso.

———. 2013. *Distant Reading*. Londres, New York: Verso.

———. [2000] 2013. « Conjectures on World Literature. » In *Distant Reading*, 43-62. Londres, New York: Verso.

———. [2006] 2013. « The End of the Beginning: A Reply to Christopher Prendergast. » In *Distant Reading*, 137-158. Londres, New York: Verso.

———. [2008] 2013. « The Novel: History and Theory. » In *Distant Reading*, 159-178. Londres, New York: Verso.

———. [2011] 2013. « Network Theory, Plot Analysis. » In *Distant Reading*, 211-240. Londres, New York: Verso.

Morse, Janice M. 1995. « The Significance of Saturation. » *Qualitative Health Research* 5 (2): 147-149. http://journals.sagepub.com/doi/pdf/10.1177/104973239500500201.

Murphy, David. 2011. « Google Abandons Street View in Germany. » *PC Mag*, 10 avril 2011. Consulté le 8 décembre 2017. https://www.pcmag.com/article2/0,2817,2383363,00.asp.

Napoli, Amedeo. 2005. A Smooth Introduction to Symbolic Methods for Knowledge Discovery. Rocquencourt (France), Le Chesnay (France): Inria. Consulté le 14 décembre 2017. https://hal.inria.fr/inria-00001210/document.

Nelson, Theodor Holm "Ted". 2014. Ted Nelson on Pernicious Computer Traditions (by Arthur Bullard). San Mateo (Californie): YouTube. Consulté le 22 novembre 2017. https://www.youtube.com/watch?v=c_KbLKm89pU.

Nunberg, Geoff. 2009. Google Books: A Metadata Train Wreck. In *Language Log*. Philadelphie: University of Pennsylvania. Consulté le 22 décembre 2017. http://languagelog.ldc.upenn.edu/nll/?p=1701.

ODCA. 2012. *Open Data Center Alliance: Big Data Consumer Guide*. Scottsdale (Arizona): Open Data Center Alliance. Consulté le 7 décembre 2017. https://bigdatawg.nist.gov/_uploadfiles/M0069_v1_7760548891.pdf.

Ohja, Ashwen. 2017. « Google Keen to Work With Gov[ernmen]t to Launch Street View in India. » *The Hindu Business Line*, 17 octobre 2017. Consulté le 6 décembre 2017. http://www.thehindubusinessline.com/info-tech/google-keen-to-work-with-govt-to-launch-street-view-in-india/article9912878.ece.

Oracle. 2013. Big Data for the Enterprise. In *Oracle White Papers*. Redwood Shores (Californie): Oracle. Consulté le 5 décembre 2017. http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf.

Oremus, Will. 2014. « The First News Report on the L.A. Earthquake Was Written by a Robot. » *Slate*, 17 mars 2014. Consulté le 12 décembre 2017. http://www.slate.com/blogs/future_tense/2014/03/17/quakebot_los_angeles_times_robot_journalist_writes_article_on_la_earthquake.html.

Pechenick, Eitan Adam, Christopher M. Danforth et Peter Sheridan Dodds. 2015. « Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. » *PLoS ONE* 10 (10). http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0137041&type=printable.

Perez, Sarah. 2017. « Twitter Launches Lower-Cost Subscription Access to its Data through New Premium APIs. » *Tech Crunch*, 14 novembre 2017. Consulté le 8 décembre 2017. https://techcrunch.com/2017/11/14/twitter-launches-lower-cost-subscription-access-to-its-data-through-new-premium-apis/.

Piazza, Alberto. 2005. « Afterword. » In *Graphs, Maps, Trees: Abstract Models for a Literary History*, sous la dir. de Franco Moretti, 95-113. Londres, New York: Verso.

Pires, Alvaro P. [1997] 2007. *De quelques enjeux épistémologiques d'une méthodologie générale pour les sciences sociales*. Chicoutimi (Québec): Université du Québec à Chicoutimi. Consulté le 6 décembre 2017. http://classiques.uqac.ca/contemporains/pires_alvaro/quelques_enjeux_epistem_sc_soc/enjeux_episte_sc_soc.pdf.

Pollock, Griselda. [1999] 2007. « Des canons et des guerres culturelles. » *Cahiers du genre* (43): 45-69. https://www-cairn-info.proxy.bibliotheques.uqam.ca:2443/revue-cahiers-du-genre-2007-2-page-45.htm.

Pustejovsky, James et Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. Pékin, Cambridge (Massachusetts), Farnham (Angleterre), Cologne (Allemagne), Sebastopol (Californie), Tokyo: O'Reilly.

Redman, Thomas C., Christopher Fox et Anany Levitin. 2003. « Data and Data Quality. » In *Encyclopedia of Library and Information Science*, 2e éd., sous la dir. de Miriam A. Drake, 782-793. New York: Marcel Dekker.

Revel, Judith. 2002. *Le Vocabulaire de Foucault*. Paris: Ellipse. Consulté le 20 juin 2014. http://1libertaire.free.fr/LeVocabulairedeFoucault.pdf.

Rheinberger, Hans-Jörg. [2006] 2010. *An Epistemology of the Concrete: Twentieth-Century History of Life*. Durham (Caroline du Nord), Londres: Duke University Press.

Ricœur, Paul. 1986. *Du texte à l'action : Essais d'herméneutique II*. Paris: Le Seuil.

Roszak, Theodore. 1986. *The Cult of Information: The Folklore of Computers and the True Art of Thinking*. New York: Pantheon Book.

Rowley, Jennifer. 2007. « The Wisdom Hierarchy: Representations of the DIKW Hierarchy. » *Journal of Information Science* 33 (2): 163-180. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.585.5962&rep=rep1&type=pdf.

Roy, Normand et Roseline Garon. 2013. « Étude comparative des logiciels d'aide à l'analyse de données qualitatives : De l'approche automatique à l'approche manuelle. » *Recherches qualitatives* 32 (1): 154-180. http://cerberus.enap.ca/ENAP/docs/Portail_etudiant/Etudiants_chercheurs/RoyGaron_2013.pdf.

Schnapp, Jeffrey, Peter Lunenfeld, Todd Presner et Johanna Drucker. 2009. *The Digital Humanities Manifesto 2.0*. Consulté le 20 novembre 2017. http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf.

Schönfelder, Walter. 2011. « CAQDAS and Qualitative Syllogism Logic: NVivo 8 and MAXQDA 10 Compared. » *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 12 (1). http://www.qualitative-research.net/index.php/fqs/article/view/1514/3134.

Schwedel, Heather. 2017. « The Pause Pod Is Genius Because You Can Crawl Inside It When People Mock You. » *Slate*, 29 septembre 2017. Consulté le 9 décembre 2017. http://www.slate.com/blogs/future_tense/2017/09/29/the_pause_pod_is_genius_because_you_can_crawl_inside_it_when_people_mock.html.

Sellars, Wilfrid. [1956] 2000. « Empiricism and the Philosophy of Mind. » In *Knowledge, Mind and the Given : Reading Wilfrid Sellars's "Empiricism and the Philosophy of Mind," Including the Complete Text of Sellars's Essay*, sous la dir. de Willem A. deVries et Timm Triplett, 205-276. Indianapolis, Cambridge (Massachusetts): Hackett.

Shannon, Claude E. [1948] 1963. « The Mathematical Theory of Communication. » In *The Mathematical Theory of Communication*, sous la dir. de Claude E. Shannon et Warren Weaver, 29-125. Urbana (Illinois): The University of Illinois Press.

Simon, Agnès. 2015. Le web de données en pratique : Data.bnf.fr In *Pour les professionnels*. Paris: Bibliothèque nationale de France. Consulté le 8 décembre 2017. http://www.bnf.fr/fr/professionnels/anx_pro_videos/a.video_cnfpt_data.html.

Simonite, Tom. 2017a. « AI Software Learns to Make AI Software. » *MIT Technology Review*, 18 janvier 2017. Consulté le 14 décembre 2017. https://www.technologyreview.com/s/603381/ai-software-learns-to-make-ai-software/.

———. 2017b. « Google's New Street View Cameras Will Help Algorithms Index The Real World. » *Wired*, 5 septembre 2017. Consulté le 8 décembre 2017. https://www.wired.com/story/googles-new-street-view-cameras-will-help-algorithms-index-the-real-world/.

Sorokin, Pitirim. [1956] 2008. *Tendances et déboires de la sociologie américaine*. Chicoutimi (Québec): université du Québec à Chicoutimi. Consulté le 16 novembre 2016. http://classiques.uqac.ca/classiques/sorokin_pitirim/tendances_socio_americaine/sorokin_tendances_socio_amer.pdf.

Spender, Dale. [1980] 1985. *Man Made Language*, 2nde éd. Londres, Boston, Sydney, Wellington: Pandora Press.

Stanford Friedman, Susan. 1991. « Weavings: Intertextuality and the (Re)Birth of the Author. » In *Influence and intertextuality in literary history*, sous la dir. de Jay Clayton et Eric Rothstein, 146-180. Madison (Wisconsin): University of Wisconsin Press. Consulté le 2 décembre 2016. http://core.roehampton.ac.uk/repository2/content2/subs/d.steedman/d.steedman2081/Stanford%20Friedman%20(1991)%20Weavings.pdf.

Statistique Canada. [2016] 2017. Aire de diffusion (AD). In *Dictionnaire, Recensement de la population, 2016 [sic]*, sous la dir. de Statistique Canada. Ottawa: Statistique Canada. Consulté le 5 décembre 2017. http://www12.statcan.gc.ca/census-recensement/2016/ref/dict/geo021-fra.cfm.

Stimson, James A., Vincent Tiberj et Cyrille Thiébaut. 2010. « Le mood, un nouvel instrument au service de l'analyse dynamique des opinions. » *Revue française de science politique* 60 (5): 901-926. http://www.cairn.info/revue-francaise-de-science-politique-2010-5-page-901.htm#re13no13.

Terra, Melissa. 2011. Peering Inside the Big Tent: Digital Humanities and the Crisis of Inclusion. In *Melissa Terra's Blog*, sous la dir. de Melissa Terra. Mountain View (Californie): Blogger. Consulté le 22 novembre 2017. http://melissaterras.blogspot.ca/2011/07/peering-inside-big-tent-digital.html.

Thérenty, Marie-Ève. 2009. « Pour une poétique historique du support. » *Romantisme* (143): 109-115. https://www.cairn.info/revue-romantisme-2009-1-page-109.htm.

Tiberj, Vincent. 2014. « Valeurs, les leçons du long terme »In *Droitisation en Europe : Enquête sur une tendance controversée*, sous la dir. de Jérôme Fourquet, Fabienne Gomant, Ernst Hillebrand et Vincent Tiberj. Paris: Fondation Jean-Jaurès. Consulté le 19 novembre 2017. https://hal-sciencespo.archives-ouvertes.fr/hal-01070790/document.

Twitter. 2017. *Search Tweets*. Twitter. Consulté le 9 décembre 2017. https://developer.twitter.com/en/docs/tweets/search/overview.

Underwood, Ted. 2013. *Why Literary Period Mattered: Historical Contrast and the Prestige of English Studies*. Stanford (Californie): Stanford University Press.

———. 2016. « The Life Cycles of Genres. » *Journal of Cultural Analytics*. http://culturalanalytics.org/2016/05/the-life-cycles-of-genres/.

Underwood, Ted et David Bamman. 2016. « The Instability of Gender. » 2016 Modern Language Association Convention, Austin Convention Center, Austin (Texas), 7-10 janvier 2016. Consulté le 25 décembre 2017. https://tedunderwood.com/2016/01/09/the-instability-of-gender/.

Uprichard, Emma. 2013. « Focus: Big Data, Little Questions? » *Discover Society*, 1er octobre 2013. Consulté le 8 décembre 2017. https://discoversociety.org/2013/10/01/focus-ig-data-little-questions/.

Vitali-Rosati, Marcello. 2015. « Paratexte numérique : La fin de la distinction entre réalité et fiction? » *Cahier ReMix* (5). http://oic.uqam.ca/fr/remix/paratexte-numerique-la-fin-de-la-distinction-entre-realite-et-fiction.

Wachowski, Lana et Lilly Wachowski. 1999. Matrix. In *Matrix*. Burbank (Californie): Warner Bros.

Warren, Virginia L. [1986]. *Guidelines for Non-Sexist Use of Language*. American Philosophical Association. Consulté le 23 décembre 2017. http://www.apaonline.org/?nonsexist.

Weaver, Warren. [1949] 1963. « Recent Contributions to the Mathematical Theory of Communication. » In *The Mathematical Theory of Communication*, sous la dir. de Claude E. Shannon et Warren Weaver, 1-28. Urbana (Illinois): The University of Illinois Press.

Wiedemann, Gregor. 2013. « Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences. » *Forum: Qualitative Sozialforschung / Forum: Qualitative Social Research* 14 (2). http://www.qualitative-research.net/index.php/fqs/article/view/1949/3552.

———. 2016. *Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany*. Wiesbaden (Allemagne): Springer.

Wilder-James, Edd. 2012. « Big Data, Big Hype: Big Deal. » *Forbes*, 31 décembre 2012. Consulté le 5 décembre 2017. https://www.forbes.com/sites/edddumbill/2012/12/31/big-data-big-hype-big-deal/#d1949c72b839.

Wilson, Mark. 2017. « AI Is Inventing Languages Humans Can't Understand: Should We Stop It? » *CoDesign*, 14 juillet 2017. Consulté le 15 décembre 2017. https://www.fastcodesign.com/90132632/ai-is-inventing-its-own-perfect-languages-should-we-let-it.

Zikopoulos, Paul C., Chris Eaton, Dirk deRoos, Thomas Deutsch et George Lapis. 2012. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York; Chicago; San Francisco; Lisbonne; Londres; Madrid; Mexico; Milan; New Delhi; San Juan (Porto Rico); Séoul; Singapour; Sydney; Toronto: McGraw Hill. Consulté le 6 décembre 2017. https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/I111025E.pdf.

Zoph, Barret et Quoc V. Le. 2017. « Neural Architecture Search with Reinforcement Learning. » 5th International Conference on Learning Representations, Palais des Congrès Neptune, Toulon (France), 23-25 avril 2017. Consulté le 13 décembre 2017. https://openreview.net/pdf?id=r1Ue8Hcxg.

Autres titres de cette collection

2015-01	**Dias da Silva**, Patricia et Lorna **Heaton** « Citizens, amateurs, volunteers: conceptual struggles in studies of citizen science »
2014-03	**Hanel**, Petr « Is China Catching up Human Health-related Applications of Biotechnology? »
2014-02	**Maroy**, C., P. **Doray**, M. **Kabore** « La politique de financement des universités au Québec à l'épreuve du « Printemps érable »»
2014-01	**Bastien**, N., P. **Chenard**, P. **Doray**, B. **Laplante** « Économie, société et éducation: l'effet des droits de scolarité sur l'accès aux études universitaires au Québec et en Ontario »
2013-03	**Hanel**, Petr, Jie **He**, Jingyan **Fu**, Jorge **Niosi** et Suzan **Reid** « A romance of the three kingdoms and the tale of two cities: the role and position of the biotechnology industry cluster in Guangdong province, China »
2013-02	**Gauthier,** Elisabeth, Gale **E. West** et Anne-Marie **Handfield** « Why do humans need to do battle? Social representations of alternative pest control approaches »
2013-01	**Bastien,** Nicolas, Pierre **Chenard,** Pierre **Doray** et Benoit **Laplante** « L'accès à l'université: le Québec est-il en retard? »
2012-01	**Prud'homme.,** Julien , Yves **Gingras**, Alain **Couillard** et Daniel **Terrasson** « Les mesures de l'interdisciplinarité. Pratiques et attitudes dans un centre de recherche français : l'IRSTEA »
2011-02	**Verdier ,** Éric, Pierre **Doray** et Jean-Guy **Prévost** « Régionalisation et recomposition du travail statistique : esquisse d'une comparaison France-Québec »
2011-01	**Mayer**, Leticia « PROBABILISM. A Cultural environment that led to the creation of random probability? »
2010-04	**Bourque**, Claude Julie, **Doray** Pierre, Christian **Bégin** et Isabelle Gourdes-**Vachon** « Le passage du secondaire au collégial et les départs des étudiants en sciences de la nature »
2010-03	**Couture**, Stéphane, Christina **Haralanova**, Sylvie **Jochems** et Serge **Proulx** « Un portrait de l'engagement pour les logiciels libres au Québec »
2010-02	**Gingras**, Yves et Sébastien **Mosbah-Natanson** « La question de la traduction en sciences sociales : Les revues françaises entre visibilité internationale et ancrage national »
2010-01	**Gingras**, Yves Naming without necessity: On the genealogy and uses of the label "historical epistemology"

CIRST

Centre interuniversitaire
de recherche sur la science
et la technologie

Le CIRST est, au Canada, le principal regroupement interdisciplinaire de chercheurs dont les travaux sont consacrés à l'étude des dimensions historiques, sociales, politiques, philosophiques et économiques de l'activité scientifique et technologique.

Nos travaux visent l'avancement des connaissances et la mise à contribution de celles-ci dans l'élaboration et la mise en œuvre des politiques ainsi que dans la résolution des problèmes de société qui présentent des dimensions scientifiques et technologiques.

Le CIRST rassemble une soixantaine de chercheurs provenant d'une douzaine d'institutions et d'autant de disciplines, telles que l'histoire, la sociologie, la science politique, la philosophie, les sciences économiques, le management et les communications.

Situé sur le campus de l'Université du Québec à Montréal (UQAM), le CIRST est reconnu comme unité de recherche par l'UQAM et l'Université de Montréal. Créé en 1986 grâce au programme des Actions structurantes du ministère de l'Éducation, le CIRST est depuis

[science, technologie, société]