

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

SCIENCE, TECHNIQUE ET SOCIÉTÉ : LE PROCÈS SOCIO-SCIENTIFIQUE
OPÉRÉ PAR LE *BIG DATA* EN SCIENCE BIO-INFORMATIQUE

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAITRISE EN COMMUNICATION

PAR
WILLIAM GRONDIN

JUIN 2020

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.10-2015). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Je souhaite accorder ma profonde reconnaissance à chacune des personnes ayant participé, de près ou de loin, à l'élaboration de cette recherche. La rédaction de ce mémoire a été rendue possible par la présence et le soutien de nombreux collègues et amis.

Un merci tout spécial à mes parents, Joëlle Bélanger et Pierre Grondin, pour le support et leurs encouragements, vous êtes mes modèles, je n'ai que de l'amour et de l'admiration pour vous. Merci de m'avoir accueilli à la demeure familiale pour le dernier mois de rédaction – quand j'étais au bout de ma vie.

Je tiens à remercier mon directeur Monsieur Maxime Ouellet pour son soutien, sa rigueur intellectuelle et les conseils avisés. Son vaste champ de connaissance est inspirant, je me sens privilégié d'avoir pu travailler avec une personne aussi érudite. Je remercie également Madame Florence Millerand et Monsieur André Mondoux d'avoir accepté de siéger au comité d'examen de ce mémoire et de m'avoir conseillé.

Au cours de mon baccalauréat en communication, politique et société à l'UQAM, j'ai rencontré Julien, Cardin, Lee, Eddy et Math qui sont par la suite devenu mes meilleurs amis. Merci pour les trop nombreuses bières, les discussions et les moments inoubliables, vous occupez une place spéciale dans ma vie. Je n'oublie pas mes collègues du *goulag post-moderne*, un vieux local de recherche sans fenêtres, habité par des bougres aussi attachants qu'in vraisemblables.

Il est impératif de mentionner le cours Études en communication : aspects épistémologiques, méthodologiques et critiques, dispensé par l'insolite duo constitué de Monsieur Éric George et Monsieur Oumar Kane. Celui-ci a particulièrement marqué mon parcours à la maîtrise, mes réflexions et élargi mes intérêts. Je remercie mes collègues du Groupe de recherche sur l'information au quotidien (GRISQ) pour les séminaires parfois trop compliqués pour 90% de l'audience, les éclats de rire et de m'avoir accueilli à bras ouvert. Je remercie également l'Association des étudiantes et étudiants à la maîtrise et au doctorat en communication de l'UQAM (AéMDC) et ses membres pour les événements, les rencontres et la vie associative.

Ces dernières années ont été parsemées de plusieurs séances de *speed* psychanalyse prodiguées par Myriam Moore et Joëlle Gélinas. Je leur suis particulièrement reconnaissant d'avoir désamorcé mes doutes et recadré mes idées.

DÉDICACE

Je dédie ce mémoire à Alexandra Asanovna Elbakyan, fondatrice de *Sci-Hub*, une base de données dont la caractéristique est de rendre disponibles, sans frais, des articles scientifiques confinés derrière des *pay-walls*. Récemment poursuivie par Elsevier, le plus grand éditeur de revue scientifique, elle se tient debout en mettant de l'avant l'idée que les connaissances scientifiques devraient être libres de droits. Les marges de profit d'Elsevier – atteignant près de 37% en 2015 – contrastent avec le coût grandissant pour obtenir l'accès au contenu des éditeurs. Le présent modèle où les membres des institutions universitaires effectuent la recherche, l'écriture des articles, le travail de révision, et ce gratuitement, pour ensuite céder les droits à des éditeurs qui revendent à prix d'or ces mêmes recherches, est défaillant et met en péril le développement social et scientifique.

TABLE DES MATIÈRES

LISTE DES FIGURES.....	viii
LISTE DES TABLEAUX.....	ix
LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES	x
RÉSUMÉ.....	xi
INTRODUCTION.....	1
CHAPITRE 1 Problématique	3
1.1 Définitions.....	5
1.1.1 Définition technique.....	5
1.1.2 Définition économique : la marchandisation des données	9
1.1.3 Un renversement dans la confiance : du gouvernement à la gouvernance	12
1.2 Science et Big Data	16
1.2.1 Contexte épistémologique : l'émergence du néo-empirisme	17
1.2.2 Portrait des Big Data en science.....	19
1.2.3 Data-driven science	21
1.2.4 Big Data et biologie moléculaire.....	22
1.3 L'Encyclopedia of DNA Elements	25
1.4 Questions de recherche	29
1.5 Perspectives.....	30
1.5.1 Perspective scientifique.....	30
1.5.2 Perspective communicationnelle.....	31

CHAPITRE 2 Cadre théorique.....	33
2.1 Déterminisme et neutralité technique.....	34
2.1.1 Du déterminisme technique.....	34
2.1.2 De la neutralité technique.....	35
2.1.3 Comment étudier un phénomène socio-technique?	36
2.2 De la société moderne à la post-modernité cybernétique.....	38
2.3 Techno-science : éléments de concrétisation	44
2.3.1 Les techno-sciences : des équipements de pointe	45
2.3.2 Les techno-sciences : légitimation stratégique.....	45
2.3.3 Les techno-sciences : concrétisation du prospectif fictionnel.....	46
CHAPITRE 3 Méthodologie.....	49
3.1 Posture du chercheur : le réalisme critique	49
3.2 Stratégie de recherche : l'étude de cas	50
3.3 La sélection du phénomène et du cas d'étude.....	51
3.4 La collecte des données : L'étude documentaire	51
3.5 Traitement et analyse des données	53
3.6 Limites et réflexions.....	54
CHAPITRE 4 Présentation des résultats.....	56
4.1 Funding Opportunity Announcement	57
4.1.1 FOA RFA-HG-16-002: Expanding the Encyclopedia of DNA Elements (ENCODE) in the Human and Mouse	59
4.1.2 FOA RFA-HG-16-003: Characterizing the Functional Elements in the Encyclopedia of DNA Elements (ENCODE) Catalog.....	59
4.1.3 FOA RFA-HG-16-004: Computational Analysis of the Encyclopedia of DNA Elements (ENCODE) Data (U01)	60
4.1.4 FOA RFA-HG-16-005: ENCODE Data Coordinating Center (U24)	61
4.1.5 FOA RFA-HG-16-006: ENCODE Data Analysis Center (U24)	61
4.2 Politique d'utilisation & de distribution des données	62

4.3 Article scientifique de référence : An Integrated Encyclopedia Of DNA Elements In The Human Genome	63
CHAPITRE 5 Discussion : de la technè à l'épistémè, de l'épistémè à la doxa	67
5.1 De la technè à l'épistémè : Big Data et science	68
5.1.1 De l'encyclopédie à la base de données : vers la mise à mort de l'encyclopaedia?.....	69
5.1.2 État du savoir : Le sujet informationnel.....	71
5.1.3 Forme de compréhension et méthode : La corrélation en tant que moyen et fin	73
5.1.4 Ambitions & conditions de possibilités	74
5.1.5 Du groupe au consortium, du chercheur au gestionnaire ;.....	75
5.2 De l'épistémè à la doxa : cybernétique et société	77
5.2.1 De la réification cybernétique	79
CONCLUSION.....	84
Réponse aux questions de recherches	86
Limites de la recherche	90
Piste de recherche : La gouvernance par les sciences de la communication? Du gouvernement à la gouvernementalité hyperréelle	91
BIBLIOGRAPHIE	94

LISTE DES FIGURES

Figure 1 Participants de la phase 4 du projet ENCODE	27
--	----

LISTE DES TABLEAUX

Tableau 1-1 Synthèse des paradigmes scientifiques selon Gray.....	20
Tableau 3-1 Présentation de la cueillette de données.....	52
Tableau 3-2 Grille thématique.....	54
Tableau 4-1 Présentation des <i>Funding Opportunity Announcement</i>	58
Tableau 4-2 Synthèse thématique de la politique d'utilisation et de distribution des données.....	63
Tableau 4-3 Synthèse thématique de l'article de référence du projet ENCODE	66

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

ADN / DNA : acide désoxyribonucléique / Deoxyribonucleic acid.

DAC : Data Analysis Center du projet ENCODE

DARPA : Defense Advanced Research Projects Agency

DCC : Data Coordinating Center du projet ENCODE

ENCODE : The Encyclopedia of DNA Elements

FOA : Funding Opportunity Announcement du projet ENCODE

HARPA : Health Advanced Research Projects Agency

NHGRI : National Human Genome Research Institute

NIH : National Institute of Health

STS : Science & Technology Studies

RÉSUMÉ

Ce mémoire consiste en une réflexion concernant l'influence du *Big Data* sur les sciences, la technique et la société. En tant que phénomène socio-technique s'étant rapidement développé, les conséquences du *Big Data* sur les sciences sont encore à définir et la littérature scientifique couvrant le sujet est incomplète. Sans plonger tête baissée dans la thèse du déterminisme technique, il paraît évident que le *Big Data* est une condition de possibilité aux effets structurants pour les sciences et la société – tout comme l'ont été l'écriture et l'imprimerie. C'est à ce titre que nous avons choisi d'étudier ce phénomène socio-technique en portant notre attention sur les sciences bio-informatiques, un domaine ayant récemment fait une place cardinale à cette technologie. Plus précisément, notre objet d'étude est *l'Encyclopedia of DNA Elements* (ENCODE), celui-ci s'inscrit comme figure de proue dans l'utilisation du *Big Data* dans ce domaine.

Notre mise en contexte s'inscrit – par le biais des notions de modernité et post-modernité – de façon à comprendre plus largement les effets des *Big Data* sur les sciences. La notion de post-modernité scientifique étant intrinsèquement liée au paradigme cybernéticien, ce travail permet de cerner l'évolution épistémologique de ce programme qui semble désormais tomber dans l'oubli, malgré ses nombreuses filiations avec les nouvelles technologies et les fondements de la post-modernité.

Pour former et analyser notre corpus, nous avons sélectionné la perspective de l'étude de cas par méthode de la recherche documentaire. Les documents sélectionnés pour analyser les tenants et aboutissants de *l'Encyclopedia of DNA Elements* se déclinent en trois types. La majeure partie de notre corpus est composée des *Funding Opportunity Announcement*, c'est-à-dire les offres de financements offertes par le *National Institutes of Health (NIH)*. On y constate que le projet est de grande envergure et organisé en consortium, c'est-à-dire que les responsabilités sont distribuées. Les documents complétant notre corpus sont la politique de distribution des données et l'article scientifique de référence du projet.

Enfin, la présentation des résultats et l'analyse closent ce mémoire en approfondissant les rapports entre technique & science, puis de la science vers la *doxa*. Les sciences étant désormais un des principaux moteurs d'idéologie, donc de formation et maintien d'un monde commun, notre projet permet de rendre compte de l'influence du *Big Data* et des sciences dans la *doxa* post-moderne, et ce, sous plusieurs points. *Primo*, que le

projet ENCODE confond le savoir encyclopédique avec les bases de données. *Secundo*, que l'Anthropos est désormais conceptualisé sous un angle strictement informationnel. *Tertio*, que la forme de compréhension relative au *Big Data* s'axe sur la corrélation en tant que moyen et fin. *Quarto*, que le *Big Data* constitue à la fois une condition de possibilité et de contingentement technique des sciences. *Quinto*, que le mode organisationnel d'ENCODE est similaire à celui d'une grande entreprise, conditionné par ses objectifs rendu possible par le *Big Data*. Nous concluons que l'épistémè cybernétique s'inscrit désormais en tant que *doxa* se répercutant dans un procès de réification du monde social et naturel sous le code binaire.

Mots-clés : *Big Data*, sciences et technique, cybernétique, bio-informatique, post-modernité, techno-science

INTRODUCTION

La *doxa* d'un *progrès* unilinéaire dans le domaine des sciences – et corolairement en tant que progrès sociétal – constitue un sujet d'étude riche et complexe. Or, le *progrès* : « n'est pas un terme neutre ; c'est un mouvement vers des fins spécifiques » (Marcuse, 1964, p. 40). Nous savons sans contredit que les sciences ne se déploient pas de manière autonome face à la société ou en tant que téléologie dénuée de contingence. Depuis les années 1980, l'étude des sciences implique nécessairement l'étude de la technique, de là l'appellation *études des sciences et de la technique* ou *science & technology studies* dans le monde anglophone (Forman, 2007).

Des nouvelles technologies favorisant le développement scientifique, le *Big Data* est certainement un moteur technique sur lequel se fondent plusieurs attentes (Kitchin et McArdle, 2016 ; Ylijoki et Porras, 2016)(Anderson, 2008 ; Ehl, 2018 ; Grant, 2012). Toutefois, en ce qui concerne des nouveautés nous apparaissant comme le seul *progrès* possible, les effets positifs sont souvent mis à l'avant-plan, au détriment des questionnements éthiques, épistémologiques et politiques. Le spectre du *progrès* accéléré relègue la compréhension du phénomène socio-technique à l'observation *a posteriori* (De Vries *et al.*, 2018 ; Latour et Biezunski, 2010).

Ce mémoire a pour but d'inverser cette tendance d'une étude d'un phénomène après coup, au moment où il est à l'état cristallisé, quasi immuable. Le *Big Bata* a déjà et aura certainement une influence marquée en science. Or, l'étudier sous l'angle du champ de recherche de l'étude des sciences et de la technique permet de jeter un regard qui rompant avec la thèse d'un seul *progrès* possible – linéaire et inévitable – en se

donnant les outils intellectuels pour mettre à nu ce phénomène complexe, prenant récemment une place marquée dans le développement scientifique.

Pour étudier ce phénomène en l'ancrant socialement et historiquement, notre cadre conceptuel explique les tenants et aboutissants de la société post-moderne – dont les fondements reposent sur le paradigme cybernétique. Notre choix méthodologique s'est arrêté sur l'étude de cas par la recherche documentaire, cette méthode permettant une analyse en profondeur d'un phénomène à partir de documents n'ayant pas nécessairement été produits à des fins scientifiques. De là découlent notre présentation des données et notre discussion qui se divise en deux parties. La première traite de la relation dialectique du *Big Data* avec la recherche scientifique. La seconde traite du passage de l'*épistémè* cybernétique à la *doxa*¹. Enfin, nous terminons ce mémoire par une conclusion portant sur les limites en plus de proposer des pistes pour des recherches futures.

¹ Grosso modo le concept d'*épistémè* fait référence à la nature du savoir scientifique, tandis que la *doxa* à l'idéologie et ce qui semble aller de soi. Pour respecter le développement logique propre à ce mémoire, les deux concepts sont expliqués plus exhaustivement dans la section discussion. On y expliquera entre autres ce qu'on entend par *épistémè* cybernétique.

CHAPITRE 1

PROBLÉMATIQUE

Les sociétés collectent informations et données depuis plusieurs millénaires, et ce, pour enregistrer et gérer leurs activités (Kitchin et Lauriault, 2014, p. 2). Nous n'avons qu'à penser à la Bibliothèque d'Alexandrie où Ptolémée II – au 2^e siècle av. J.-C. – avait l'intention d'y stocker un exemplaire de chaque écrit pour faire de cette bibliothèque la plus grande source de savoir du monde. On estime de nos jours que la quantité d'information présente par habitant est équivalente à plus de 320 Bibliothèques d'Alexandrie (Mayer-Schönberger et Cukier, 2014, p. 25).

La révolution Gutenberg vers la fin du 15^e siècle a permis des changements significatifs quant à la production, l'organisation et la diffusion des connaissances, de la littérature et du savoir scientifique. L'imprimerie est d'ailleurs un facteur clé rendant possible la naissance et la généralisation des principes caractérisant les sciences modernes tel que nous les connaissons aujourd'hui (Huff, 2003). En permettant de mettre de l'avant les principes de transmission écrite à grande échelle, la généralisation de l'écrit face à l'oral, la publication de revue scientifique révisée par les pairs, la technique de l'imprimerie a structuré les principes de transmission de la science moderne (Eisenstein, 2005). Autrement dit, l'imprimerie est une condition de possibilité du développement de la science moderne – tel qu'elle s'est effectuée – en permettant la diffusion d'ouvrages et de publications de nature scientifique. La dissémination des ouvrages scientifiques s'est développée de façon exponentielle, tandis que la standardisation typographique a permis une reproduction plus exacte et une reproduction des figures et graphiques plus fidèles (Logan, 2004).

Les changements énoncés précédemment concernant l'imprimerie se sont déroulés sur une période de plusieurs siècles. De son côté, la *révolution* informatique et le phénomène du *Big Data* ont des effets marqués sur les sociétés, et ce, sur une période de quelques décennies, voire années. Outre des changements concernant les moyens de transmettre et de structurer l'information, le *Big Data* est un artéfact technique mobilisé dans le domaine scientifique faisant émerger des considérations nouvelles, par exemple sur le plan éthique et épistémologique.

Cette recherche a pour objectif d'étudier les contributions du *Big Data* en science à l'aune de la société post-moderne. Le *Big Data* – au-delà d'une importante quantité de données – est un phénomène ayant des répercussions techniques, politiques, économique, épistémologique et scientifique (Kitchin, 2014 ; Leonelli, 2014 ; Ruppert *et al.*, 2017). Les retentissements dans le domaine des sciences sont encore en développement, les discussions entourant le sujet sont loin d'être fixées. Une catégorisation du phénomène sous l'angle de la post-modernité nous permet de cadrer le phénomène et de contextualiser ses effets en tenant compte de ses antécédents historiques.

Pour débiter, nous présentons une définition du *Big Data* sous les angles techniques, économiques et relatifs à l'épistémologie. Par la suite, notre revue de littérature porte sur la question du *Big Data* en science. Nous y abordons la question épistémologique sous l'angle de l'approche néo-empiriste et présentons une critique de cette théorie. En dernier lieu, nous présenterons brièvement le cas du projet ENCODE, un super projet s'imposant dans le domaine de la bio-informatique.

1.1 Définitions

Comme mentionné précédemment, nous tenons à contextualiser le *Big Data* sous plusieurs aspects afin de clarifier adéquatement ce phénomène complexe. Pour ce faire, nous traitons initialement des aspects techniques du phénomène *Big Data*. Ensuite, nous abordons l'aspect économique de la nouvelle valeur des données par dépossession, ce que plusieurs nomment l'économie numérique ou informationnelle. La contextualisation se poursuit avec l'aspect de la nouvelle confiance envers les données et la gouvernamentalité algorithmique.

1.1.1 Définition technique

Cette section se penche spécifiquement sur les aspects techniques caractérisant le *Big Data*. Plus précisément, nous présentons le phénomène sous l'angle d'un renversement technique en tant que condition de possibilité. Les dimensions techniques abordées dans cette section doivent être comprise en tant que composition, c'est-à-dire qu'elles sont constituantes du *Big Data* dans la perspective d'une logique d'ensemble, opérant en synergie pour former un tout cohérent.

Nous abordons la composition des 3V – volume, vitesse, variété – du *Big Data*. Les articles *What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets* de Kitchin & McArdle (2016), *Perspectives to Definition of Big Data: A Mapping Study and Discussion* de Ylijoki & Porras (2016), *Empirical Big Data Research: A Systematic Literature Mapping* de Mathisen & al (2015) et l'article *Big Data : A review* par Sagioglu & Sinanc (2013) permettent de jeter un méta-regard sur la définition du concept.

Pour ce groupe d'auteurs, bien que le concept de *Big Data* puisse être défini de plusieurs façons, c'est le volume, la vitesse et la variété qui le caractérise avant tout. Cette conclusion est tirée de leurs études dont un des buts est de définir les

caractéristiques du phénomène en effectuant des méta-études en analysant la littérature scientifique sur le sujet. Ainsi, nous justifions la pertinence de présenter ces 3 aspects à partir des études citées précédemment.

1.1.1.1 Volume : vers un monde numérique

Au début des années 2000, uniquement le quart de l'information stockée l'était sous format numérique, la majorité étant disponible exclusivement sous format analogique, comme des ouvrages imprimés ou des cassettes magnétiques (Mayer-Schönberger et Cukier, 2014, p. 24). En une période de 7 ans – soit moins d'une décennie – un important renversement quant à l'analogique et le numérique se produit. L'information disponible et produite passe majoritairement de l'analogique vers le numérique. On estime qu'en 2007, c'est seulement 7% de l'information qui est disponible exclusivement sous format analogique, le reste l'étant sous format numérique (Mayer-Schönberger et Cukier, 2014, p. 24).

Cette tendance s'accroît pour atteindre seulement 2% des données disponibles sous format analogique en 2013, comparativement à 98% pour le format numérique. Ce renversement s'explique par l'expansion rapide de la production des données numériques, celles-ci doublant environ tous les 3 ans, tandis que la production de données analogique demeure marginale face à l'avalanche numérique (Mayer-Schönberger et Cukier, 2014, p. 25).

Par exemple, ce sont 5 exabytes de données qui ont été créées par l'Anthropos en date de 2003 (Sagiroglu et Sinanc, 2013, p. 41). Une dizaine d'années plus tard, ce sont environ 2.5 exabytes qui sont produits quotidiennement et ce nombre devrait doubler par période de plus ou moins 40 mois (McAfee *et al.*, 2012, p. 4). Au total, ce sont 8 zetabytes de données présentes en 2015 contre seulement 5 exabytes en 2003. On estime que ce sont environ 50 milliards d'appareils qui seront connectés à des réseaux internet d'ici 2020 et qui produiront du même coup une importante quantité de données.

La tendance est donc que de plus en plus d'appareils *intelligents* produisant des données sont mis en service et que chacun d'entre eux a une capacité de génération accrue comparativement à la précédente.

Le volume de données stockées, en circulation et analysées, nécessite une considération spéciale et dépasse la capacité d'un simple ordinateur. Comme l'explique Kitchin et McArdle, c'est souvent en *terabyte*, *petabyte*, voire en *exabyte* que les bases de données se chiffrent². Le volume – sous format numérique – est donc un des volets constitutifs des 3 V du *Big Data*. Bien que la quantité de données à elle seule ne puisse constituer le *Big Data*, elle en est certainement un élément clé de celui-ci (Gordon, 2013, p. 12).

1.1.1.2 *Vélocité : création et analyse en temps réel*

Plus qu'un volume important de données, le *Big Data* implique une capacité de captation et de traitement quasi instantanée : « A conventional understanding of velocity typically considers how quickly the data is arriving and stored, and its associated rates of retrieval. » (Zikopoulos, 2012, p. 8). Au-delà de cette définition, c'est le : « [...] data in motion : The speed at which the data is flowing », le flux de donnée constant et massif qui caractérise la vélocité (Zikopoulos, 2012, p. 8). Le traitement et l'analyse des données ne se comptent plus en minutes ou en secondes, mais en millisecondes. Puisque les données sont mises à jour en temps réel, les analyses sont faites pendant que les données sont en mouvement – donc en temps réel – plutôt qu'à l'état statique. La vélocité s'inscrit dans un régime de temporalité quantitatif – en opposition à un régime qualitatif – où l'objectif est l'atteinte de l'instantanéité. Il n'y aurait plus de distance entre le passé et le futur, seulement un régime temporel présentiste où le temps immédiat s'actualise perpétuellement, déconnecté d'une quelconque historicité.

² Un *byte* équivaut à un bloc de 8 *bits* composé de 1 ou 0. Un *terabyte* est composé de 10^{12} *bytes*, un *petabyte* de 10^{15} *bytes* et un *exabyte* de 10^{18} *bytes*.

Pour Kitchin et McArdle, la vitesse est la caractéristique la plus significative pour définir le *Big Data* (Kitchin et McArdle, 2016, p. 7). Ils dénotent deux définitions convergentes de la vitesse, soit la vitesse de la fréquence de génération et la vitesse de traitement et de publication. Autrement dit, le processus de génération et le traitement de l'information s'effectuent de façons simultanées et sans relâche. L'information demeure en perpétuel mouvement autant dans sa captation, sa génération, son stockage et son analyse. Ce renversement contraste avec la conception d'une captation et d'un traitement des données qui – il n'y a pas si longtemps – était chronophage et complexe.

1.1.1.3 Variété : hétérogénéité compatible

Le concept de variété appliqué au *Big Data* est la capacité de traiter simultanément des données hétérogènes provenant de plusieurs sources. La variété est le traitement et l'analyse croisée de plusieurs bases de données stockant des données diversifiées autant dans le fond qu'au niveau de la forme (Mathisen *et al.*, 2015, p. 5). En connectant plusieurs bases de données, des corrélations globales ou par sous-groupes se dessinent en plus de pouvoir former des corrélations multi-variables (Gray *et al.*, 2015, p. 2).

En ajoutant une importante variété dans les sources et dans les données elles-mêmes, en conjonction avec un volume important relativement modeste, le traitement des données permet d'ajouter des couches de complexités analytiques favorisant l'analyse par croisements multiples. La corrélation entre deux variables demeure importante pour faire état d'un phénomène, mais il est désormais possible de déceler des corrélations entre des centaines de variables de sources différentes.

En plus d'inclure des données provenant de bases de données multiples, la variété dans le traitement s'inscrit dans l'hétérogénéité structurelle des données qui peuvent être structurées, semi-structurées et non-structurées. Les données structurées sont représentées par des données déjà classées et accessibles, tandis que les données non-

structurées sont difficilement accessibles et classables. Corolairement, la variété est également présente dans les :

[...] analysis methods as distributed programming, pattern recognition, data mining, natural language processing, sentiment analysis, statistical and visual analysis and human computer interaction. Therefore architecture must support various methods and analysis techniques (Sagiroglu et Sinanc, 2013, p. 47).

En résumé, la variété s’inscrit dans la diversité des variables, des sources, des méthodes et du type de donnée pour arriver à effectuer une analyse en produisant des corrélations et des résultats.

1.1.2 Définition économique : la marchandisation des données

Le discours économique dominant – concernant la multiplication des données – soutient que celles-ci sont le pétrole de demain. Or, un discours plus critique remet en perspective la marchandisation de la vie quotidienne par dépossession et accumulation des données (Thatcher *et al.*, 2015, p. 1). Le *Big Data* permet des transformations économiques importantes en étant une condition de possibilité pour que les données deviennent une marchandise à grande échelle. Comme l’explique Thatcher :

[...] we parse a variety of Big Data definitions to argue that it is only when individual datums by the million, billion or more, are linked together algorithmically that ‘Big Data’ emerges as a commodity. Such decisions do not occur in a vacuum but as part of an asymmetric power relationship in which individuals are dispossessed of the data they generate in their day-to-day lives. We argue that the asymmetry of this data capture process is a means of capitalist “accumulation by dispossession” that colonizes and commodifies everyday life in ways previously impossible (Thatcher *et al.*, 2015, p. 1-2).

La marchandisation des données à travers le *Big Data* est le prolongement attendu du système économique capitaliste qui tend à intégrer et récupérer – de façon hégémonique – les espaces concrets et abstraits en s’adaptant et en structurant les façons

de faire et la technique. La *datafication*³ du monde se produit de façon asymétrique au sens où les individus produisent une quantité importante d'information et de donnée qui est ensuite stockée, échangée et au bout du compte marchandisée.

Cette accumulation par dépossession est caractéristique du phénomène *Big Data* que nous connaissons aujourd'hui et fait écho à la nouvelle valeur des données. Par processus de dépossession, nous entendons ici un état de dissolution systématique des communs par un capitalisme communicationnel⁴ décuplant les processus d'appropriation des données et les données elles-mêmes (Thatcher *et al.*, 2015). Les données sont dorénavant concentrées entre les mains de grandes corporations comme Google, Amazon, Facebook, Apple, Microsoft (GAFAM) – pour fin d'accumulation – en incluant les sujets producteurs pour aussitôt les déposséder de leurs propres données (Dean, 2016, p. 4).

Comparativement à une dépossession capitaliste concrète – où par exemple un lieu devient physiquement inaccessible – la dépossession des données est abstraite, parfois imperceptible. Cette dépossession se produit en deux temps, une marchandisation continue du *moment* présent en convertissant chaque geste anodin en données, puis en limitant le contrôle des utilisateurs quant à leurs données. Le *futur* n'est pas en reste, le passé et le présent étant stocké et analysé dans un mode prédictif, le capitalisme communicationnel médié par le *Big Data* recherche des modèles pour prédire et aligner le futur vers l'objectif de faire disparaître toute forme de contingence (Dean, 2016, p. 5). Dès lors, la génération d'une quantité importante de données est nécessaire pour atteindre cet objectif de prédiction s'apparentant à l'omniscience. La norme est dorénavant de produire le plus de données possible en orientant les dispositifs techniques, les lois et les façons de faire vers cet objectif (Thatcher *et al.*, 2015, p. 3).

³ La *datafication* est la tendance à socio-technique à numériser l'ensemble des aspects de la vie (Matteson, 2018).

⁴ Par capitalisme communicationnel ou informationnel, on entend un régime de valorisation économique axé sur la captation et l'agrégation des données.

Nous identifions 3 principaux aspects caractérisant la marchandisation des données dans une *société de l'information*⁵. En premier lieu, l'émergence du *Big Data* orienté vers le marché de l'information dans le cadre d'une croissance infinie. En deuxième lieu, l'extension de la vie privée comme terrain d'accumulation et de dépossession médié par les technologies de l'information. En troisième lieu, la relation asymétrique dans la production et l'extraction des données est soutenue par la quantification et un éminent appareillage de surveillance (Thatcher *et al.*, 2015, p. 3). Ces aspects convergent vers la quantification totale de la vie humaine en tant que capital.

Le processus de transformation des données en marchandise fait partie d'un ensemble économique de production et d'extraction de la valeur des données. La tendance capitaliste de colonisation d'un nouveau champ ou espace se caractérise souvent par un type d'accumulation par dépossession et le présent cas n'en fait pas exception (Alvarez León, 2016 ; Harvey, 2006 ; Thatcher *et al.*, 2015).

Ce processus de privatisation des données les concentre entre les mains des entreprises du secteur technologique et isole les données individuelles pour les quantifier, transformant du même coup les données et les individus en abstraction marchande. Concrètement, c'est une perte de contrôle des données personnelles en utilisant des services des médias socionumériques. Les données qui semblaient autrefois *discrètes* et peu pertinentes sont transformées dans un processus abstrait en pièces aux multiples dimensions. Il y a inversement dans la fonction des appareils et des services. Si autrefois la captation des données était secondaire à un dispositif technique ou un service, cela devient dorénavant la fonction première et vice-versa (Goodchild, 2007, p. 217).

⁵ Le terme *société de l'information* fait référence à un discours à la fois médiatique, politique et scientifique dont le fondement est la potentialité d'une nouvelle société basée sur les technologies de l'information et de la communication. Ce discours à saveur techno-déterministe propose une société où l'information et les données sont le principal moteur économique et social. Le déploiement massif des nouvelles technologies de la communication est alors vu comme inévitable et nécessaire à la bonne gouvernance (George et Granjon, 2008).

Le processus de dépossession est particulièrement patent dans le cas des réseaux sociaux numériques où les utilisateurs acceptent un contrat à travers les conditions d'utilisations souvent longues, caractérisées par un vocabulaire juridique complexe⁶. Par exemple, les conditions d'utilisation de Facebook stipulent que les données personnelles produites par les utilisateurs peuvent être échangées et monétisées par le géant de plusieurs manières, mais ces règlements sont en constant changement pour englober encore plus de manières d'y arriver (Smith, 2013).

La marchandisation des données ouvre la porte à la création d'un marché des données personnelles des consommateurs. À titre d'exemple, la firme Acxiom – qui se spécialise dans la vente et l'échange de données personnelles – a dégagé un revenu de plus de 1 milliard en 2011. Cette firme dispose de plus de 23 000 serveurs et analyse plus de 50 billions de transactions concernant plus de 200 millions de personnes par année. La base de données de ce courtier en données comprend en moyenne 1500 points d'informations par personne, incluant : « [the] age, race, sex, weight, height, marital status, education level, politics, buying habits, household health concerns, vacation dreams, etc. » (Roderick, 2014, p. 729).

1.1.3 Un renversement dans la confiance : du gouvernement à la gouvernance

La confiance est centrale dans la construction et le maintien de l'ordre social (Shapin, 2007). Elle agit comme ciment sociétal favorisant la formation d'un monde commun. La production du savoir s'assoit sur un ensemble complexe d'interactions matérielles et sociales prenant pour acquis la fiabilité et la stabilité de nombreuses normes et institutions (Rieder et Simon, 2016, p. 3). On note que les systèmes de confiance sont variables dans l'espace et dans le temps et qu'ils sont sujets à des modifications conjoncturelles ou structurelles.

⁶ Sur une note à la fois humoristique et empreinte de légèreté, nous précisons qu'il faudrait environ 76 journées de travail par années pour lire exhaustivement l'ensemble des conditions d'utilisation acceptées au cours d'une période de 12 mois (Wagstaff, 2012).

Bernard Cohen dans l'ouvrage *The triumph of numbers: how counting shaped modern life* (2005) – faisant la généalogie historique des nombres – avance qu'une des caractéristiques fondamentales de la modernité est l'usage prépondérant des chiffres dans la majorité de la socialité (Cohen, 2005, p. 12). En effectuant une rigoureuse histoire des chiffres – en prenant soin d'effectuer une revue de littérature des exemples les plus mobilisés et pertinent dans le champ de l'histoire des sciences – Cohen soutient que la mobilisation des chiffres remonte à des milliers d'années, mais qu'elle est toutefois peu comparable à la forme qu'elle prend aujourd'hui. C'est précisément dans ce système que la quantification du monde relative au *Big Data* prend place, et ce, dans une relation coextensive. Cette nouvelle forme de confiance s'axe sur les chiffres et les données plutôt qu'envers les institutions elles-mêmes (Rieder et Simon, 2016, p. 3).

Dans son ouvrage *Trust in numbers : the pursuit of objectivity in science and public life* (1995), Theodore Porter soutient que la gestion s'effectue dorénavant par les chiffres et que la confiance envers les institutions ne suffit plus, qu'il est nécessaire de mettre en place un régime d'objectivité « mécanique » basé sur les nombres et les données (Porter, 1995, p. 90). Cette forme de gouvernementalité⁷ met l'accent sur la réduction de la présence humaine pour favoriser une approche axée sur des principes formalisés et standardisés où la subjectivité humaine serait absente (Daston et Galison, 1992, p. 82 ; Rieder et Simon, 2016, p. 4). Autrement dit, le renversement de la confiance s'opère entre une confiance basée sur des institutions et des individus, vers une confiance envers : « [...] highly formalized procedures enacted through disciplined self-restraint. Numbers cease to be supplements. They are couched in a rhetoric of factuality, imbued with an ethos of neutrality, and presented with an aura of certainty » (Rieder et Simon, 2016, p. 4).

⁷ La gouvernementalité est un concept – dont on attribue l'origine à Foucault – portant sur le procès de rationalité propre à la gouvernance d'une société moderne, et ce, croisé à une perte de confiance envers les institutions. Dès lors, la gouvernementalité algorithmique ou politique publique basée sur les données probantes est une force prescriptive ayant pour objectif d'encadrer l'administration étatique dont le point focal est les données (Lafleur, 2017).

Rieder et Simon (2016) soutiennent dans l'article *Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data* que les nouvelles technologies permettent un prolongement de ce changement dans les modes de confiance. Les ordinateurs sont le symbole ultime de l'opposition entre humain et non-humain, la dichotomie caractéristique de l'impureté et de la subjectivité face à la machine bienveillante et neutre qui veille en permanence (Venturini *et al.*, 2014, p. 6).

Un exemple concernant l'importance de la confiance dans nos sociétés est observable dans le domaine des instituts de la statistique. Les auteurs Peter Struijs, Barteld Braaksma et Piet Daas (2014) dans l'article *Official statistics and Big Data* avancent que dans le cadre d'une société post-moderne les statistiques occupent une place prépondérante dans la mise en place de politiques publiques. Cependant, les statistiques peuvent uniquement participer à la mise en place des politiques publiques basée sur les données probantes : « [...] if they can be trusted. In advanced societies, official statistics are often taken for granted, but where trust is lacking, society misses an important pillar for informed discussion and evidence-based policy-making » (Struijs *et al.*, 2014, p. 1).

Sous cet angle, les institutions adoptent de façon généralisée un mode de gestion processuel des données, au sens où la confiance envers celles-ci doit passer par des standards et des procédures, voire par l'imaginaire d'un effacement de la partialité humaine. Le processus de gouvernance par les nombres agit comme tremplin de la confiance et le *Big Data* comme catalyseur de ce processus. (Struijs *et al.*, 2014, p. 3). Ce changement concernant les modes de confiance n'est pas le simple résultat du hasard, mais reflète la perte de confiance généralisée dans les institutions politiques. D'ailleurs, Lyotard annonce plus largement ces changements dans le cadre de la fin des grands récits de la modernité. La perte de créance des grands récits ou méta-récits, dont les institutions modernes s'étaient harnachées pour légitimer leurs existences, fait en sorte que le monopole narratif de l'émancipation bouleverse les fondements institutionnels (Pagès, 2011).

Les récentes crises économiques autant que les crises démocratiques laissent leurs marques. Confrontée à méfiance publique, la confiance envers les institutions se transporte du côté de la confiance envers la gouvernance par les nombres, propulsée par le *progrès* électronique et l'informatisation du monde (Rieder et Simon, 2016, p. 4). Conséquemment, où la méfiance prévaut, la confiance envers des processus stricts et la quantification du monde comblent un espace laissé vacant :

It is on these grounds that the Big Data phenomenon continues to blossom [...] Big Data connect to the ideal of mechanical objectivity in several ways, not only fortifying but also expanding the appeal of the doctrine : [...] Big Data pledges to extend the reach of automation, from data collection to matters of storage, curation, and analysis. The virtuous machine emerges as ever more powerful as it covers increasingly large parts of the analytical and decision-making process (Rieder et Simon, 2016, p. 4).

En résumé, le *Big Data* s'inscrit dans un contexte de crise de la confiance institutionnelle et plus largement une crise politique, une perte de sens dissolvant les liens sociaux modernes. La quantification du monde se positionne en tant que comme nouvelle modalité de formation de sens et de repère, puisque les précédents seraient devenus caducs.

C'est le concept de *gouvernementalité algorithmique* qui résume le plus justement la transition vers la gouvernance axée sur la processualité et les systèmes informatiques carburant aux données massives (Rouvroy et Berns, 2013 ; Rouvroy et Stiegler, 2015). L'état de gouvernementalité algorithmique passe par « trois temps » qui, en se combinant de manière calculée, forment un régime de gouvernementalité basé sur la processualité et l'algorithmisation de la socialité.

Le premier temps est la collecte, le classement et l'accumulation automatisés de vastes quantités de données, et ce, de façon systématique. Dans le cas de la gouvernementalité algorithmique, la provenance des données est reliée à une multitude de sources, puis conservée dans des centres de données dont la taille est virtuellement illimitée. On note que la collecte et la conservation des données ne sont pas nécessairement reliées à des

impératifs de gouvernance immédiats, mais à l'accumulation en ayant foi que les données seront plus tard mobilisées en vue de n'importe quel objectif pragmatique (Rouvroy et Berns, 2013, p. 169).

Le deuxième temps est le traitement systématique et instantané des données collectées, ce qu'on nomme *datamining*. Le but du traitement est de faire ressortir un ensemble de corrélations à partir de données dont la classification est hétérogène et décontextualisée. L'intervention humaine doit être tenue au minimum, les hypothèses initiales sont mises de côté pour écarter le potentiel de subjectivité, donc d'un savoir dont l'essence serait l'objectivité (Rouvroy et Berns, 2013, p. 170).

Le troisième et dernier temps est la concrétisation de l'usage des données dans un modèle statistique d'anticipation des choix et des comportements individuels. C'est l'application pratique et normative d'une logique statistique d'anticipation des attentes et actions individuelle, qu'elles soient bien réelles ou potentielles (Rouvroy et Berns, 2013, p. 172).

Ces trois temps se confondent en se justifiant dans une logique d'ensemble dont la cohérence est catalysée par la coextensivité et la mutualité. Toujours dans cette logique, par gouvernementalité algorithmique nous désignons :

Un certain type de rationalité (a)normative ou (a)politique reposant sur la récolte, l'agrégation et l'analyse automatisée de données en quantité massive de manière à modéliser, anticiper et affecter par avance les comportements possibles (Rouvroy et Berns, 2013, p. 173).

1.2 Science et *Big Data*

La section précédente a permis de dresser un portrait du *Big Data* quant aux aspects techniques, économiques et relatifs à la gouvernementalité. Or, le *Big Data* a également une incidence prononcée sur le domaine scientifique. La présente section aborde

premièrement le contexte épistémologique d'émergence du *Big Data*, celui-ci étant marqué par les prémices du néo-empirisme. En second lieu, nous dressons un portrait de l'évolution des paradigmes scientifiques en soulignant les changements s'opérant sous le spectre des données massives. En troisième lieu, on explique le modèle des *data-driven sciences*, celui-ci met de l'avant un modèle inductif dont les données sont l'étincelle d'un processus de recherche. En dernier lieu, on précise notre sujet d'étude en abordant la thématique de la bio-informatique.

1.2.1 Contexte épistémologique : l'émergence du néo-empirisme

Le célèbre article de Chris Anderson *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* publié en 2008 dans *The Wired* a donné lieu à une littérature foisonnante et est particulièrement cité dans l'étude des questions épistémologiques relatives au *Big Data*. La thèse de l'auteur projette que l'avalanche de données et la possibilité technique de les traiter instantanément ouvrent la porte pour faire des études scientifiques sans mobiliser de théories (Anderson, 2008). Toujours selon Anderson, les données sont tellement massives qu'elles ne peuvent plus être simplement visualisées, il faut les traiter mathématiquement et les contextualiser ultérieurement.

En mobilisant des algorithmes, des systèmes informatiques complexes et une quantité importante de données, les sciences humaines comme la sociologie et la psychologie deviendraient caduques au sens où les données pourraient désormais parler d'elles-mêmes (Anderson, 2008). Sous cet angle, les données massives oblitéreraient la nécessité d'utiliser un modèle ou une théorie, il serait même nuisible d'en tenir compte. L'analyse des données se ferait sans préjugé ou idée préconçue, des *patterns* statistiques seraient mis à jour, ce que la science traditionnelle ne pourrait pas accomplir. Plutôt que de créer ou mobiliser des modèles explicatifs, on recherche des corrélations entre des variables dans un but prédictif. La vision d'Anderson s'inscrit dans un courant néo-empiriste et technophile où la science *Big Data* permettrait d'accéder à une vérité qui était autrefois hors de portée.

La théorie néo-empiriste concernant le *Big Data* est marquée par plusieurs postulats. Le premier est qu'il serait possible de pouvoir cerner l'ensemble des données disponibles concernant un phénomène, l'échantillonnage serait désuet et inefficace face à la potentialité de capter l'ensemble des données. Deuxièmement, la mobilisation théorique serait négative au sens où les données pourraient désormais parler d'elles-mêmes. En mobilisant des théories, on empêcherait les données de pleinement révéler un phénomène. Troisièmement, puisque les données peuvent parler d'elles-mêmes, elles seraient non-humaines, il faudrait donc écarter le plus possible les manipulations humaines pour atteindre l'objectivité. Quatrièmement, il n'y aurait plus de contexte et de discipline spécifique, le savoir serait ouvert à quiconque sait traiter les données. Ces quatre caractéristiques suggèrent un nouveau mode scientifique de nature purement inductif (Kitchin, 2014, p. 4). La rationalité des sciences modernes, qui consiste à lier un phénomène à une cause, est contraire à la rationalité scientifique post-moderne dont l'objectif est de prévoir au détriment du comprendre (Rouvroy et Stiegler, 2015, p. 109).

La vision néo-empiriste d'Anderson est très loin de faire l'unanimité et a donné lieu à de vives discussions critiques. La question épistémologique est loin d'être évacuée et constitue une portion importante des discussions sur la place du *Big Data* en science. Étant donné que la thèse d'Anderson est l'archétype de la position technophile et qu'elle est souvent prise en contre-exemple pour ensuite en faire une critique, il nous semble judicieux de maintenant présenter une critique de celle-ci.

Les critiques avancent que les postulats néo-empiristes sont simplement erronés. Si l'idée de pouvoir capter l'entièreté des données relatives à un phénomène semble intéressante, elle ne pourrait être plus fallacieuse. La captation des données est à la fois régulée par la capacité technique de récolte et par des *a priori* concernant le choix des données qui sont récoltées (Crawford, 2013). Or, penser pouvoir capter et traiter l'ensemble des données d'un phénomène revient à rechercher une certaine omniscience. Les données ne sont donc pas naturelles ou des abstractions extraites

d'une réalité objective et neutre (Kitchin, 2014, p. 5). Deuxièmement, les algorithmes et les méthodes d'analyses n'émanent pas de nulle part. La méthode inductive mise de l'avant dans ce genre d'analyse est cadrée discursivement par des recherches précédentes. La neutralité et l'objectivité des algorithmes est au mieux une illusion, au pire un discours occultant le caractère social des données. Troisièmement, les corrélations ne parlent pas d'elles-mêmes. Concrètement, elles doivent être identifiées et interprétées pour faire du sens, c'est-à-dire que des corrélations peuvent n'avoir aucun lien causal. Il faut donc se garder de laisser les données parler d'elles-mêmes, il est important d'effectuer un rigoureux travail de vérification et de comprendre les corrélations. Quatrièmement, l'idée que les disciplines s'effacent et que quiconque accompagné de matériel technique et d'une certaine connaissance statistique puisse interpréter les données est simplement fallacieuse. Les différents domaines scientifiques emploient un jargon, des concepts, des théories et des objets d'étude les caractérisant. Il ne suffit pas d'avoir des connaissances statistiques pour être psychologue, sociologue ou chercheur en physique. L'interprétation des données implique nécessairement des connaissances que ce soit pour cerner des corrélations fallacieuses ou bien pour comprendre des phénomènes (Kitchin, 2014, p. 5).

1.2.2 Portrait des Big Data en science

Le *Big Data* n'est pas uniquement un phénomène technique consistant en une grande quantité de données. Ceci est particulièrement patent dans le domaine scientifique où la dimension symbolique des données s'apparente à celle de l'économie informationnelle, c'est-à-dire que les données auraient désormais une valeur attribuable à un processus d'accaparement (Leonelli, 2014, p. 2). En acquérant le statut de *marchandise* scientifique, les données sont des extrants reconnus à la fois dans le domaine scientifique et à l'extérieur de celui-ci. Également, le *Big Data* en science inscrit de nouvelles méthodes, façon de procéder, de gestion d'information et de collaboration (Leonelli, 2014, p. 2).

Bien que les promesses du *Big Data* en science soient parfois exagérées, plusieurs discours coexistent. Certains avancent une véritable révolution scientifique, tandis que d'autres sont plus modérés. Par exemple, dans l'article *What difference does quantity make? On the epistemology of Big Data in biology*, Leonelli (2014) avance que : « [...] the natural sciences may well be the area that is least affected by Big Data, whose emergence is much more likely to affect the political and economic realms – though not necessarily for the better. » (Leonelli, 2014, p. 2).

Faisant écho à la théorie des paradigmes scientifiques de Kuhn où des : « découvertes scientifiques universellement reconnues qui, pour un temps, fournissent à une communauté de chercheurs des problèmes types et des solutions » (Kuhn, 1983, p. 11), Gray avance que ce même processus se produit également de façon générale en science (Hey, 2009, p. xvii ; Kitchin, 2014, p. 3). Ce processus, au même titre qu'un paradigme kuhnien, se produit de façons successives comme une nouvelle extension de la méthode scientifique, mais potentiellement dans l'ensemble des domaines scientifiques.

Tableau 1-1 Synthèse des paradigmes scientifiques selon Gray

Paradigme	Nature	Forme	Période
Premier	Science expérimentale	Empirisme : description de phénomène naturel	Pré renaissance
Deuxième	Science théorique	Modélisation et généralisation	Pré ordinateur
Troisième	Science computationnelle	Simulation de phénomène complexe	Pré <i>Big Data</i>
Quatrième	Science exploratoire (<i>Big Data</i>)	Basé sur les données, les statistiques et le <i>data-mining</i>	Maintenant

On note que selon Gray, les changements paradigmatiques en sciences sont relatifs à des façons de faire et leur forme est reliée à une période historique. Les plus récents

changements sont principalement propulsés par les avancées technologiques dans le domaine de l'électronique et l'informatique. (Tableau traduit et adapté de (Kitching, 2014), lui-même repris de (Hey et al, 2009).

1.2.3 *Data-driven science*

Ce sont les termes *data-driven science*, *data-intensive science* et *big science* qui traduisent le mieux la mobilisation du *Big Data* et d'une méthode scientifique impliquant des modèles et théories. Rappelons que la méthode néo-empiriste est basée sur 4 postulats.

- Un procès de captation des données quasi omnisciente ;
- Un procès d'analyse uniquement inductif ;
- Des corrélations qui parlent d'elle même ;
- L'effacement des disciplines au profit de la statistique comme procès de compréhension absolue ;

Nous abordons indirectement et directement ces postulats dans la présente section en comparant la posture du néo-empirisme développée par Anderson et l'approche *data-driven science*. Contrairement au discours néo-empiriste, le modèle *data-driven science* n'avance pas le postulat de laisser les données parler d'elle-même sans mobiliser des concepts et de théories :

In contrast to new forms of empiricism, data-driven science seeks to hold to the tenets of the scientific method, but is more open to using a hybrid combination of abductive, inductive and deductive approaches to advance the understanding of a phenomenon (Kitchin, 2014, p. 5).

La méthode *data-driven science* propose un mode initialement inductif où la finalité n'est cependant pas une explication émanant uniquement des données comme l'approche néo-empiriste. De plus, l'induction ne sort pas de nulle part, elle est théoriquement située, contextualisée et peut employer des modalités déductives : « As

such, the epistemological strategy adopted within data-driven science is to use guided knowledge discovery techniques to identify potential questions (hypotheses) worthy of further examination and testing » (Kitchin, 2014, p. 5).

Dans ce processus scientifique, la théorie existante agit à titre de guide orientant les recherches, au lieu de simplement rechercher des corrélations entre des données sans aspirer à expliquer théoriquement une corrélation. Mieux, la théorie pourrait servir de « garde-fou » pour être critique face à la signification de ces corrélations. Les données sont récoltées de façon ciblée relativement à un ancrage théorique fort, contrairement à la méthode néo-empiriste qui cherche à récolter le plus de données possible pour espérer révéler une vérité invisible, qui se trouverait pourtant sous nos yeux (Kitchin, 2014, p. 5). Les modalités analytiques ne sont pas strictement statistiques – voire disciplinairement et théoriquement décontextualisées – elles s’inscrivent dans un domaine et comprennent des concepts spécifiques à des sujets d’étude.

Le processus abductif permet en partie de rendre compte de la méthode des *data-driven sciences* au sens où : « Abductive reasoning is a form of inference that starts with data describing something and ends with a hypothesis that best explains the data » (Miller, 2010, p. 194). Le modèle abductif est particulièrement important dans le processus scientifique impliquant le *Big Data*, les données massives sont prises comme étincelles pour entamer un processus de découverte scientifique. Par exemple, en étudiant un ensemble de données et en relevant des corrélations, il est possible d’entamer une recherche scientifique en mobilisant des théories et concepts pour comprendre et ensuite expliquer le phénomène.

1.2.4 Big Data et biologie moléculaire

Contrairement à l’article *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* mentionné précédemment, où Anderson indique que la science n’aura plus besoin de modèle, concept et théorie, l’étude des sciences biologiques

moléculaires montre que *Big Data* et modèle scientifique peuvent très bien cohabiter. La prédiction d'Anderson s'est minimalement partiellement réalisée, au sens où le *Big Data* – aussi nommé *big science* dans le domaine scientifique – est dorénavant bien présent dans les plusieurs disciplines sciences, mais coexiste avec les modèles et théories (Leonelli, 2014 ; Marx, 2013).

La multiplication de la puissance des ordinateurs, de pair avec la diminution des coûts de stockage des données, permet aux laboratoires plus modestes de devenir des générateurs d'une grande quantité de données (Marx, 2013). Ironiquement, les données générées prennent numériquement et physiquement plus d'espace que les instruments mobilisés pour les créer (Marx, 2013). Cet enjeu demeure relativement récent, mais s'est développé plus rapidement qu'autrement. S'il y a à peine une dizaine d'années le domaine des sciences bio-moléculaires pouvait aisément maintenir un rapport de proximité avec les données, il produit désormais des données à un rythme plus qu'important : « Compared to fields like physics, astronomy and computer science that have been dealing with the challenges of massive datasets for decades, the Big Data revolution in biology has also been quick, leaving little time to adapt » (Singer, 2013).

En une courte période temporelle, on passe d'un modèle où il est possible d'extraire la pleine valeur des données, à une fulgurante production de données nécessitant des considérations techniques spécifiques où l'extraction de la valeur est complexe et requiert toujours plus de données. Cette transition est d'ailleurs explicitement reconnue par le directeur du *National Human Genome Research Institute*, mentionnant au passage que les sciences de la vie : « [...] are becoming a Big Data enterprise » (Singer, 2013).

Les projets de type *Big Data* en biologie moléculaire sont majoritairement virtualisés et focalisés sur le *cloud computing*⁸. L'entièreté des données est habituellement stockée dans des centres de données, tandis que les laboratoires n'en stockent localement qu'une fraction (Marx, 2013). Le *cloud computing* est particulièrement populaire dans la recherche scientifique en biologie moléculaire, dans la mesure où les subventions peuvent se faire rares, les laboratoires n'ont pas nécessairement à participer financièrement au maintien de l'infrastructure physique pour avoir accès aux données (Marx, 2013). Or, certains vont jusqu'à faire l'analyse directement sur des serveurs externes. Au-delà du stockage de données, le *cloud* permet de traiter informatiquement les données à distance. L'analyse des données s'effectue avec la puissance computationnelle des centres de données, ce qui est un avantage pour les laboratoires limités par des contraintes budgétaires (Marx, 2013).

En ce sens, les grands projets scientifiques en science bio-moléculaire ont une importante fonction de production, de classement et de diffusion des données, mais leur fonction analytique est possiblement limitée. Puisque les données des grands projets sont souvent disponibles gratuitement, la voie est pavée pour que les projets d'envergure moindre puissent analyser ou se servir des données : « The role of the Big Data projects is to sketch the outlines of the map, which then enables researchers on smaller-scale projects to go where they need to go » (Singer, 2013).

L'ancien rédacteur en chef de la revue *Science* témoigne de la cohabitation entre les projets importants en science bio-moléculaire et de leur relation avec des laboratoires d'envergure limités :

Ensuring a successful future for the biological sciences will require restraint in the growth of large centers and -omics—like projects, so as to provide more financial support for the critical work of innovative small laboratories striving to understand the wonderful complexity of living systems » (Alberts, 2012).

⁸ Le *cloud computing* – infonuagique en français – est l'utilisation de serveur informatique à distance pour effectuer des calculs ou stocker de l'information (Armbrust *et al.*, 2010)

Selon lui, une fois que les projets d'envergures ont produit des données, le travail analytique devrait en partie revenir aux plus petits laboratoires considérant les implications théoriques complexes.

1.3 L'Encyclopedia of DNA Elements

Les sciences bio-moléculaires sont un domaine qui n'a été que récemment touché par la vague *Big Data*, bien que l'informatisation de la bio-génétique ne date pas d'hier. Il y a à peine une dizaine d'années, les données générées étaient d'envergure moindre et ne représentaient pas un enjeu logistique et analytique complexe : « the field was relatively data poor, and scientists could easily keep up with the data they generated » (Singer, 2013). L'Encyclopedia of DNA Elements est la concrétisation d'un projet scientifique de type *Big Data* en science bio-moléculaire et est sans contredit la référence d'une entreprise scientifique d'envergure dans ce domaine.

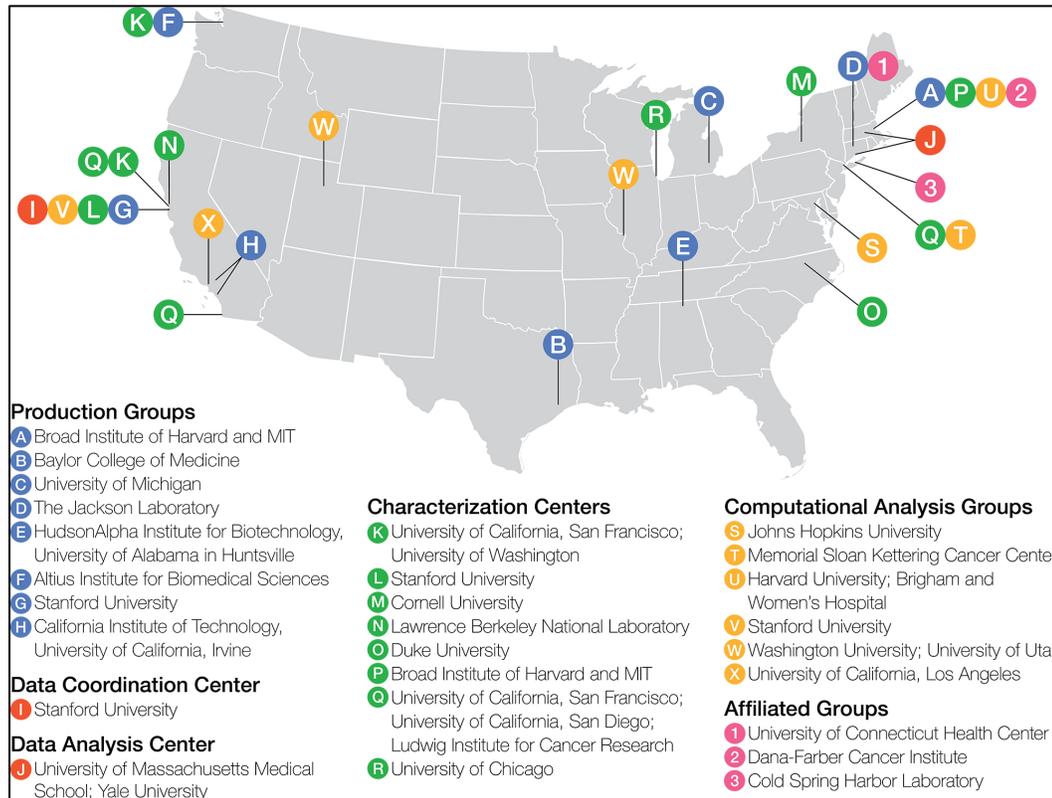
Officiellement instauré en 2007, soit dans une période temporelle correspondant aux débuts du *Big Data*, il est une figure de proue d'une méthode de travail en ce qui concerne les *big sciences* dans le domaine bio-moléculaire. Le projet ENCODE a débuté sous sa forme pilote en 2003 dans le but de continuer les démarches du *Human Genome Project*, il est entré dans sa phase de production en 2007-2012. De 2012 à 2016, c'est la troisième phase du projet qui s'amorce, notamment en étendant son champ de recherche vers des organismes non-humains (ENCODE, 2014). En 2017 le projet est entré dans sa quatrième phase de financement, cette phase ajoute : « [...] a way to incorporate data provided by the research community, and will use biological samples from research participants who have explicitly consented for unrestricted sharing of their genomic data » (National Human Genome Research Institute, 2017).

Dans sa phase pilote de 2003 à 2007, le projet a porté son attention sur l'analyse de seulement 1% du génome humain. Bien que des doutes importants concernant la

capacité du projet de s'étendre, les conclusions de la partie pilote du projet ont eu des retombées importantes. Avant la publication des résultats du Projet ENCODE, on avançait que 80% de l'ADN humain était sans fonction, qu'elle était du *junk DNA*. Or, la partie pilote du projet a démontré que cela était potentiellement faux, que même si une partie de l'ADN n'était pas une protéine encodante, ses parties avaient d'autres fonctions significatives. Cette percée majeure au stade initial du projet a ouvert des portes dans le monde de la génétique, par exemple en montrant que :

DNA variants previously correlated with certain diseases lie within or very near non-coding functional DNA elements, providing new leads for linking genetic variation and disease. [...] These findings force a rethink of the definition of a gene and of the minimum unit of heredity (Ecker *et al.*, 2012, p. 52).

L'objectif global du projet est de dresser la liste la plus complète possible des éléments du génome humain. Ces éléments, agencés dans des proportions adéquates : « constitute the information needed to build all the types of cells, body organs and, ultimately, an entire person from a single genome » (Ecker *et al.*, 2012, p. 52). Compte tenu de la complexité et de la taille du projet, celui-ci ne peut pas fonctionner avec une petite équipe d'un ou deux laboratoires. Au contraire, le projet est organisé en consortium, c'est-à-dire en groupe ayant un objectif commun, incluant plus de 30 laboratoires et environ 450 scientifiques provenant d'horizons disciplinaires divers (Birney, 2012, p. 49).



(ENCODE Project Consortium, 2019b)

Figure 1-1 Organisation de la phase 4 du projet ENCODE

Le projet *ENCODE* est généralement reçu positivement – certains vont même jusqu’à le qualifier de véritable : « Google Maps of the human genome » (Kolata, 2012). D’autres, comme le professeur Dan Graur de l’Université de Houston au Texas, critique ouvertement ENCODE pour son peu de rigueur intellectuelle, un montant alloué d’environ 300 millions depuis le début du projet et des opérations de relation publique pour s’assurer d’obtenir des subventions (Graur *et al.*, 2013)

Selon lui, bien que la génération et le classement de données effectué par le consortium sont exemplaires, l'analyse qui y est faite demeure au mieux faible, au pire fallacieuse (Graur *et al.*, 2013). Il avance que les projets de type *big science*, comme le projet ENCODE, devraient seulement produire des données tandis que les projets de type *small science* devraient effectuer l'analyse. Selon le critique : « Encode authors are computational scientists, not biological scientists. " They're computer jocks," as Graur put it. "Big science, like the Human Genome Project, should publish data » (Boyle, 2013).

En ce qui concerne la mobilisation du *Big Data* et l'organisation du projet. ENCODE mobilise :

[...] unprecedented amounts of data, [...] present new computational and data-analysis challenges and have been a major force driving the development of computational methods in genomics. [...] These processing pipelines and quality-control measures have been adapted by the research community as the standard for the analysis of such data (Ecker *et al.*, 2012, p. 55).

Autrement dit, ce projet de recherche se base principalement sur une grande quantité de données synthétisées à l'aide de méthodes normalisées entre les différents participants du projet. De plus, le consortium effectue un travail de vérification des données pour s'assurer qu'elles sont de qualité et reflète les standards de production du projet. À maintes reprises, il a été qualifié de projet de type *big science* par plusieurs scientifiques et est souvent repris comme l'archétype de ce type de projet (Alberts, 2012 ; Birney, 2012 ; Boyle, 2013 ; Ecker *et al.*, 2012).

La politique de donnée ouverte du projet fait en sorte qu'il joue un rôle critique dans l'accessibilité aux données bio-informatiques : « [...] to be used as a reference for the scientific community » (Birney, 2012, p. 49). La totalité des données générées est disponible sur le site web du consortium et peut à la fois être recherchée, visualisée en ligne et téléchargée (ENCODE, 2018). En 2013, on estime que le projet totalisait plus de 18 téraoctets de données brutes (Singer, 2013). En utilisant une approche normalisée

à travers les différents laboratoires générant des données, *ENCODE* peut créer des catalogues et des ressources importantes plutôt que de créer un petit nombre de points d'intérêt.

1.4 Questions de recherche

Compte tenu des éléments présentés dans notre problématique, nous constatons que la place du *Big Data* en science n'est pas cristallisée. À la fois artéfact et processus techniques, le *Big Data* opère des changements significatifs dans maintes disciplines scientifiques, autant dans le mode organisationnel, le financement, voire jusque dans le régime épistémologique des sciences. Certains vont même avancer que c'est carrément un changement paradigmatique qui s'opère dans le monde scientifique – changement dont l'origine serait le *Big Data*. Cela nous mène à la question de recherche suivante :

De quelle façon le *Big Data* influence-t-il les sciences et la société ?

De notre question générale, deux questions spécifiques émergent.

Comment le *Big Data* influence l'épistémè scientifique ?

De quelle façon les sciences *Big Data* s'inscrivent dans la *doxa*?

1.5 Perspectives

1.5.1 Perspective scientifique

À la suite de notre revue de littérature, nous soulignons que plusieurs réflexions quant au *Big Data* sont encore au niveau liminaire. Ceci est particulièrement patent dans le domaine des sciences *naturelles* où les théories critiques sont absentes ou théoriquement faibles face à ce que plusieurs traitent comme une véritable révolution des données en science (Charter *et al.*, 2007).

Notre apport scientifique s'inscrit dans le prolongement des recherches en cours dans le domaine des *data studies*. De nombreuses recherches en sciences sociales s'intéressent au phénomène du *Big Data*, en transitant entre plusieurs champs disciplinaires ou objets de recherche. La posture que nous proposons, c'est-à-dire d'étudier le *Big Data* en science, permet de poursuivre les recherches existantes et de rendre compte du processus d'informatisation et de technicisation de la science contemporaine. Bien que des questionnements et des critiques épistémologiques concernant le *Big Data* aient concrètement été émis, ils demeurent fondés sur des objets analysés de façon partielle et dont la contextualisation plus large dans le domaine épistémologique de l'étude des sciences demeure limitée (Canali, 2016 ; Kitchin, 2014 ; Lowrie, 2017 ; Rieder et Simon, 2016).

Le potentiel des *Big Data* en science et les transformations déjà établies par ce changement technique ne sont plus à débattre. Or, ces transformations ne sont peut-être qu'à leur balbutiement, n'étant réellement en place – pour des raisons principalement techniques – que depuis une décennie (Charter *et al.*, 2007). Plus précisément, notre étude est originale dans la mesure où elle contextualise le *Big Data* en science sous des catégorisations socio-historiques en mettant l'accent sur l'évolution et les changements épistémologiques s'opérant dans le domaine scientifique. De plus, en effectuant une étude de cas sur l'emblématique projet ENCODE, les théories portant sur le *Big Data*

en science sont mises en relation avec un cas ancré empiriquement. Cet élément permet de répondre à notre question de recherche en fournissant une critique sur le plan épistémologique, contribuant à la compréhension de ce phénomène.

1.5.2 Perspective communicationnelle

La perspective communicationnelle de notre projet d'étude du *Big Data* fait d'abord écho à l'essence des sciences et de la technique. Le processus scientifique, que ce soit la conservation, la production et le partage du savoir sont liés aux sciences de l'information et de la communication (Garvey, 1979). Récemment, un changement plus profond attire notre attention, l'épistémologie scientifique s'est muée en épistémologie de la communication. L'épistémologie des sciences est synonyme de communication dans la mesure où le programme cybernéticien transcende le domaine scientifique (Sfez, 1990). La fusion entre la communication et l'épistémologie scientifique – dont le *Big Data* est une matérialisation parmi d'autres – confond le savoir et la connaissance au profit d'une vision strictement informationnelle du monde social et naturel (Lafontaine, 2004). Étudier le *Big Data* en science, c'est donc porter un regard sur l'algorithmisation du monde converti en données binaires.

Le monde naturel et social fusionne dans l'*épistémè* informationnelle pour être converti en corpus de données s'alignant sur une forme et un code commun. La subjectivité du social s'efface au profit du langage binaire, la nature est quant à elle traduite vers la racine commune de l'informatique. L'hégémonie informatique fait en sorte que le seul savoir valable, voire réel, est celui qui se présentera sous forme binaire (Lyotard, 1979). Le *Big Data* en science est la matérialisation de l'effacement des frontières classiques entre le vivant et le non-vivant, au profit d'une compréhension algorithmique où : « La réduction de l'humain à une série d'informations complexes, [est] la principale conséquence de l'emprise du modèle informationnel sur les sciences de la vie » (Lafontaine, 2004, p. 206).

Les sciences bio-informatiques, de concert avec le *Big Data*, représentent ce changement épistémologique dans la compréhension du savoir. Si autrefois les sciences de la vie proposaient une réflexion plus large sur l'expérience du sujet vivant, la bio-informatique réduit le sujet humain à un code informatique. La frontière entre corps humain et machine s'efface au profit d'un dénominateur commun, le code génétique stockable sur disque dur. Ce changement épistémologique en science est de nature communicationnelle et va de pair avec la reproduction symbolique du monde fondé sur la médiatisation effective du réel. Plus précisément, c'est la réification scientifique du monde social et naturel sous le paradigme cybernétique qui constitue le point d'ancrage communicationnel de ce mémoire.

CHAPITRE 2

CADRE THÉORIQUE

Le deuxième chapitre constitue l’ancrage théorique de notre mémoire. Il est divisé en 3 sections s’inscrivant dans une logique d’ensemble, contextualisant les aspects pertinents dans le cadre de notre recherche portant sur les *Big Data* et le développement des sciences. À l’instar d’Andrew Feenberg, nous concevons que l’étude d’un phénomène socio-technique nécessite de s’inscrire dans une approche incluant une théorie englobante concernant la société où la technique se dénoue. De surcroît, il est nécessaire de tenir compte de la dialectique technique-société pour comprendre l’influence de la technique sur la société et vice-versa (Feenberg *et al.*, 2014, p. 255).

La première partie présente les concepts de déterminisme et de neutralité de la technique. La présentation de ces deux concepts est nécessaire pour contextualiser le développement des sciences et de la technique, en plus de lier leur développement à la société. *Grosso modo*, le courant déterministe avance que le développement technique est téléologique – que la société devrait corolairement se mouler à ce développement qui s’inscrit inévitablement vers des stades supérieurs. La théorie de la neutralité technique propose de concevoir la technique sans essence propre – en tant que reflet de la socialité.

La seconde section précise ce que nous entendons lorsqu’on aborde la thématique de la société. Pour démontrer le contexte social dans lequel la technique se développe, nous expliquons les catégories globales de modernité et de post-modernité. Comme la post-modernité n’est pas un concept fini, il est plus que pertinent d’expliquer sa relation

à la modernité pour l'intelligibiliser et comprendre le processus de formation des sciences post-modernes. On y traite également des fondements du paradigme cybernétique et de ses liens de filiation avec la post-modernité.

La dernière section s'intéresse aux diverses matérialisations que prennent les technologies. On y aborde notamment le sujet des équipements à la fine pointe, le processus de légitimation stratégique et de la concrétisation de la fiction technique en réalité concrète.

2.1 Déterminisme et neutralité technique

Cette première section de notre cadre théorique s'inscrit comme fil d'Ariane à nos considérations d'études. Nous tenons bien humblement à dépasser certains préjugés qui selon nous, engendrent une compréhension au mieux parcellaire, au pire erronée, du phénomène *Big Data* en science, voire dans l'étude des sciences et de la technique en général. Nous y traitons sommairement de plusieurs dichotomies que nous considérons comme essentielles à réconcilier pour étudier un phénomène socio-technique. On souhaite dépasser le débat caustique entre déterminisme et neutralité technique en proposant d'être prudent face à une science qui serait en constante *evolutio*⁹.

2.1.1 Du déterminisme technique

La thèse déterministe est fondée sur la désocialisation des sciences et de la technique. Elle avance que la technique se développerait intégralement hors de la socialité, de façon autonome ayant une logique fonctionnelle propre : « [...] sans référence à la société » (Feenberg *et al.*, 2014, p. 44). Dans cet ordre d'idée, le développement des

⁹ En latin classique – donc bien avant la théorie de l'évolution de Darwin – évolution (*evolutio*) est entendu au sens de déroulement. La finalité d'une *evolutio* serait au mieux connue d'avance, au pire inconnue, mais inévitable.

sciences et de la technique se ferait selon des objectifs autonomes, en dehors de contraintes structurelles ou conjoncturelles. C'est en quelque sorte une technique réifiée, sans généalogie historique que nous propose la thèse déterministe. Sous cet angle, ce n'est pas la société qui fonde la technique, c'est la technique en tant que force exogène qui jette son influence la société (Feenberg *et al.*, 2014, p. 45).

Toujours selon la position déterministe, le progrès scientifique et technique prend une forme unilinéaire fondée sur deux affirmations que nous considérons nécessaire de questionner. *Primo*, le progrès avance nécessairement dans une seule direction possible, telle une *evolutio*, c'est-à-dire avec une fin propre et prédéterminée vers des stades nécessairement plus élevés. *Deuxio*, le développement se fait nécessairement en suivant des stades selon une série de succession logique (Feenberg *et al.*, 2014, p. 46). Si nos institutions sociales n'ont pas d'influence sur la technique, alors c'est la relation inverse qui s'instaure, c'est-à-dire que ce sont les institutions sociales qui doivent être plastiques, se mouler à l'influence technique.

Pour entamer une étude probante sur le sujet des sciences et de la technique, une saine méfiance doit s'instaurer face à la thèse déterministe. Les sciences et la technique sont – plus souvent qu'autrement – considérées comme le pôle de neutralité et de progrès par excellence. Le discours techno-capitaliste entourant le développement fusionne science et économie dans une tirade en faveur du seul progrès possible. Or, ce discours teinté de déterminisme technique ne doit surtout pas être pris à la lettre si l'on souhaite sincèrement entamer un processus d'étude des sciences et de la technique. S'il est vrai que les sciences et la technique affectent la société, on ne doit pas perdre de vue que le développement techno-scientifique est le produit de celle-ci.

2.1.2 De la neutralité technique

D'un autre côté, certains avancent que la technique serait neutre et singulièrement plastique. Une des plus célèbres tirades concernant la supposée neutralité technique est

à propos des armes à feu : « Ce ne sont pas les armes à feu qui tuent les gens, ce sont les gens avec des armes à feu ». La technique n'aurait plus d'essence, mais serait simplement le reflet de la socialité. Dépossédée d'une quelconque essence, elle se prêterait aussi bien à une administration autoritaire : « [...] mais, dans un contexte social différent, elle pourrait très bien être gérée démocratiquement » (Feenberg *et al.*, 2014, p. 41). Autrement dit, les sciences et la technique porteraient des valeurs et seraient un corolaire infléchi à l'idéologie dominante (Caidi *et al.*, 2005, p. 3). Bien que nous considérons que la technique est nécessairement liée à la socialité, nous partageons les mises en garde de Feenberg concernant ce positionnement :

[...] les formes modernes d'hégémonie sont fondées sur un type spécifique de médiation technique des différentes activités sociales, que ce soit la production ou la médecine, l'éducation ou l'armée et, par conséquent, la démocratisation exige des changements radicaux autant techniques que politiques. (Feenberg *et al.*, 2014, p. 41)

En ce sens, il ne suffit pas de démocratiser les institutions politiques pour changer la technique, il faut également démocratiser la technique elle-même pour espérer transformer le politique. Un exemple souvent mentionné pour exprimer cette dimension est le système technique de production en Union des républiques socialistes soviétiques, qui était en compétition directe avec le système capitaliste américain. L'appareillage technique, l'organisation du milieu de travail et les impératifs de production n'étant pas si différents d'un bloc à l'autre.

2.1.3 Comment étudier un phénomène socio-technique?

L'avènement du *Big Data* est rendu possible en raison d'une structure à la fois technique, scientifique, politique et économique. Au-delà d'un simple résultat relié à des conditions de possibilités, le *Big Data* est en relation coextensive d'influence mutuelle avec les structures sociales et la technique. Cette relation constitue d'après nous un aspect fondamental à prendre en compte lors de l'étude du phénomène.

D'un côté, étudier le *Big Data* comme un phénomène scientifique naturel ou relatif uniquement au développement de l'informatique de pointe, en omettant volontairement ou involontairement son caractère social, revient à participer à la réification du phénomène en le déconnectant historiquement et socialement. De l'autre, étudier le phénomène simplement comme un résultat conjoncturel social revient à passer sous silence les vastes transformations opérées par celui-ci.

Cette posture d'étude du *Big Data* fait plus largement écho à la nécessité d'aborder les phénomènes de façon pluridisciplinaire en mettant de côté la distinction classique entre phénomènes sociaux et phénomènes naturels qui caractérisent la modernité (Feenberg, 2013, p. 106 ; Latour, 2005). Notre but ici reste bien humble, sans tenter d'offrir une alternative complètement nouvelle à cette distinction moderne, nous proposons simplement de tenir compte de l'hybridation de ces phénomènes.

La majorité des études concernant le *Big Data* sont a-critiques et focalisées vers le *marketing*, l'informatique et la gestion (Kitchin, 2014, p. 3). De surcroît, le discours techno-économique ambiant occulte la dimension sociale de la technique en la réifiant. Ce discours est insidieux dans la mesure où l'opposition aux thèses techno-déterministes reste minimale, y compris dans le domaine scientifique. Les questionnements par rapport à l'utilisation des objets techniques dans le domaine scientifique sont balayés par cette foi aveugle en la technique. Le nouvel équipement, toujours plus performant et précis, permettrait simplement de faire « de la meilleure science ».

Pour étudier le *Big Data* il faut donc considérer le phénomène : « by paying attention not only to the mechanical, but also to the mental workings of an otherwise opaque phenomenon » (Rieder et Simon, 2016, p. 2). Autrement dit, nous soutenons une posture épistémologique où il est impossible, pour étudier le *Big Data*, de l'extraire de son historicité et de sa socialité en le considérant uniquement comme un phénomène technique a-historique, déconnecté du monde social. Nous proposons de concevoir le

Big Data sans reproduire les banalités du discours techno-économique ou d'une *doxa* viciée par une foi aveugle envers le progrès scientifique et technique. La technique n'est pas déterminante ou déterminée, elle est sociale tout en s'inscrivant dans une médiation, elle n'est pas simplement le résultat d'un filtrat idéologique.

Un grand travail de Marx a d'ailleurs été, dès la moitié du 19^e siècle, de repolitiser l'économie en s'opposant à l'idée d'un champ économique réifié qui serait gouverné par des lois naturelles. La dernière moitié du 20^e siècle n'a pas été moins riche concernant le sujet technique (Feenberg *et al.*, 2014, p. 39). Nous avançons qu'au 21^e siècle, à l'ère d'une société de l'information, une des grandes tâches en science sociale est de prolonger et de faire murir ces questionnements. Nous souhaitons modestement que ce projet de mémoire participe au processus de resocialisation des sciences et de la technique sous son angle socio-communicationnel.

2.2 De la société moderne à la post-modernité cybernétique

Dans le cadre de cette section, nous proposons d'étudier la post-modernité en l'examinant face au passage de la tradition à la société moderne. On y constate que la post-modernité s'inscrit comme glissement face à la modernité. Avant de nous intéresser directement à la post-modernité, nous proposons de brèves explications d'ordre conceptuelles pour comprendre la transition de la société traditionnelle vers la modernité. L'idée de modernité sous-entend une cassure entre deux modes de reproduction sociale. C'est dans cet ordre d'idée que Freitag utilise le terme de modernité pour :

désigner le concept idéal-typique qui peut être dégagé d'une analyse sociologique du statut formel d'un nouveau *mode de reproduction sociale* qui a été instituée, de manière globalement systématique et cumulative, pendant la période historique que les historiens désignent depuis le XVIII^e siècle comme celle des " Temps modernes " » (Freitag, 2002, p. 64)

Le mode de reproduction sociétale moderne, qualifié de politico-institutionnel par Freitag, est caractérisé par l'institutionnalisation du pouvoir. Dans cette logique d'ensemble, l'institutionnalisation se concrétise à travers l'État de droit constitutionnel dans le pouvoir législatif et la figure de la citoyenneté libre dans la mesure de la loi. Évidemment, sans nécessairement intelligibiliser la modernité sous ces seuls aspects, ceux-ci sont particulièrement importants considérant leur extension aux divers domaines de la socialité.

La modernité s'inscrit dans une opposition concrète et réfléchie aux anciennes structures traditionnelles. Bien qu'initialement parcellaires, les manifestations d'opposition entre des institutions, des classes sociales et l'Église se sont généralisées et formalisées dans des changements pérennes et concrets. De plus, la naissance de la science moderne entre le 16^e et le 18^e siècle s'inscrit dans une dimension coextensive d'influence mutuelle dans la possibilité et la concrétisation de la modernité telle que nous la connaissons. Dans une logique d'ensemble, ces changements convergent vers :

Une nouvelle conception cohérente de la personne humaine et de son autonomie, et que celle-ci [...] en est venue à contester de front l'ensemble des autorités traditionnelles (Freitag, 2002, p. 70).

La naissance de la modernité s'inscrit donc dans une cumulativité historique concrète. Elle n'est pas une observation abstraite *a posteriori*, elle est une logique d'ensemble opérationnelle concrétisant une opposition explicite envers la société traditionnelle. Du même coup, l'opposition de la modernité à la tradition permet de l'objectiver en la cantonnant au passé. C'est dans son rapport réel et concret d'opposition entre tradition et modernité que la société traditionnelle est définie dans l'altérité, reléguée dans un rapport temporel linéaire où elle fait partie du passé. On notera qu'une partie de la justification concernant l'impérialisme et le colonialisme s'explique justement par le supposé archaïsme des sociétés traditionnelles, donnant ainsi tous les droits aux modernes.

Enracinée dans la Raison, la question du pouvoir dans les sociétés modernes occupe une place particulière et la distingue de la tradition. Contrairement aux sociétés traditionnelles où le pouvoir s'inscrit dans un régime ontologique transcendantal, la modernité offre une alternative mettant de l'avant un idéal républicain. En premier lieu, le pouvoir étatique est dorénavant défini par : « la *capacité législative* et non plus seulement par les capacités exécutives et juridictionnelles » (Freitag, 2002, p. 74). Le principe d'autorité légitime entre dans un processus réflexif sur l'ordonnement de la socialité et la gestion du monde commun. Deuxièmement, le gouvernement obtient sa légitimité au regard de maintien du principe d'État de droit, soit la dépendance de l'État face aux institutions juridiques et législatives. Troisièmement, le processus électoral est le lien de légitimité permettant à l'assemblée de prendre des décisions (Freitag, 2002, p. 75). Dans le même ordre d'idée, l'organisation et le fonctionnement de l'État moderne – dans sa concrétisation de reproduction politico-institutionnelle – se produisent en tant qu'effectivité positive externe. L'ordre sociétal repose sur le respect des institutions, l'épanouissement de l'individualité s'exprime au conditionnel du respect des lois.

En ce qui a trait à la relation entre la modernité et la post-modernité, elles ne sont pas caractérisées par l'opposition concrète distinguant la société traditionnelle et la modernité. Bien que la modernité et la post-modernité contrastent dans leur effectivité, leur rapport n'est pas strictement oppositionnel. Si la tradition et la modernité étaient engagées dans un combat frontal, alors la modernité et la post-modernité sont engagées dans une opération se produisant dans le noyau moderne, se propageant suivant une logique d'ensemble plus subtile (Freitag, 2002, p. 74). Autrement dit, la post-modernité est entendue en tant que dissolution progressive – partielle ou totale – de la modernité.

Comme processus de substitution progressif, la post-modernité effectue une parcellisation de la socialité au détriment d'un idéal commun manifesté dans l'État en tant que pouvoir légitime unique représentant la totalité du corps social. Plus précisément, le mode de reproduction décisionnel-opérationnel se concrétise par le

passage de l'universalité vers le présentiel parcellaire axé vers l'efficacité (Freitag, 2002, p. 77).

Sur le plan épistémique, la modernité propose une séparation stricte entre l'objet et le sujet, entre le vivant et le non-vivant et de l'individu et de la société. En mettant de l'avant ce dualisme hypostasié, les sciences modernes proposent corolairement 3 postulats d'ensemble ; ontologique, méthodologique et épistémologique (Freitag, 2002, p. 97). Ontologiquement, les phénomènes sont réductibles à une nature universelle et totale. Méthodologiquement, c'est par l'expérience empirique et méthodique que la Nature se découvre dans une optique de connaissance. Du côté épistémologique, les mathématiques sont universelles au sens où toutes relations quantifiables peuvent s'exprimer à travers des variables et calculs. Toujours dans une logique d'ensemble, ces 3 postulats confluent vers un cadre scientifique caractéristique de la modernité. Les phénomènes naturels peuvent se réduire, mais font sens dans une totalité universelle, proposant de surcroît l'idée d'une Nature pouvant et devant être totalement découverte. La finalité de la science moderne demeure toutefois cognitive et axée vers la connaissance plutôt que la subordination effective en abstraction de la Nature et de la Société comme dans la post-modernité techno-scientifique (Freitag, 2002, p. 98).

Les termes : science post-moderne, techno-science et cybernétique ont en commun de traiter d'un même sujet en des termes qui peuvent paraître initialement différents, mais qui convergent dans le fond comme dans la forme (Lafontaine, 2004). Ils ont en commun de constater d'importants changements dans la nature de l'activité scientifique, phagocytée par l'idéologie de la communication. La post-modernité implique un changement radical dans l'activité scientifique, la science et la technique sont désormais régulées par la rationalité de l'efficacité, par de nouvelles normes

épistémologiques et par des visées nouvelles concernant les buts de la connaissance (Clain, 1989).

Jean-François Lyotard dans son ouvrage *La condition postmoderne: rapport sur le savoir* (1979), explique l'émergence de la nouvelle condition du savoir. L'épistémologie cybernéticienne entame sa domination des sciences dès les années 50, l'informatisation générale de la société fait en sorte que le langage binaire devient le méta-langage universel. Autrement dit, la condition de validité du savoir est son acquisition, classement et exploitation dans le méta-langage binaire (Lyotard, 1979). Cette mise en commun – autant du monde vivant et non-vivant – que du monde naturel et social, fait en sorte d'effacer les anciennes frontières modernes au profit d'un commun médié par les techniques de la communication.

Ce glissement épistémologique vers une cybernétisation totale de la société renvoie par ailleurs à des impératifs quant à la compréhension du savoir, de son origine à sa finalité. La manière d'interroger et de répondre doit être composée dans le méta-langage binaire pour passer le test de la condition de validité. Autrement dit, la réponse importe peu, tant qu'elle se dévoile dans le langage et la méthode désirée par l'hégémonie cybernéticienne.

On retrace les origines du paradigme cybernétique à la fin des années 40 dans les écrits du mathématicien Norbert Wiener. Celui-ci est considéré comme le père de ce courant, dont un des but était de rassembler des domaines comme les mathématiques, la philosophie, la médecine et la physique vers un programme commun de compréhension du monde (Wiener et Rosenblueth, 1946). Au cours de ses recherches au National Institute of Cardiology of Mexico – portant sur les caractéristiques de l'appareil cardiaque et du système nerveux – Wiener note d'importantes similarités entre les systèmes nerveux et le langage binaire (Wiener, 1950).

À première vue anodines, ces recherches vont cristalliser le programme cybernéticien dans une perspective unifiant biologie et mathématique. Pour Wiener, le corps humain et celui des animaux peuvent désormais être explicitement comparés à des machines. Dès la fin des années 1940, les théories du programme cybernétique passent conceptuellement du système synaptique des vivants à des systèmes de communications électroniques. Par exemple, une synapse et un appareil de communication binaire fonctionneraient de manière semblable : « [...] la mémoire chez l'animal trouve son parallèle dans le problème de la construction de mémoires artificielles pour les machines » (Wiener, 2014 [1948], p. 75). Cette mise en commun du vivant et du non-vivant est la norme plutôt que l'exception dans l'œuvre du mathématicien. C'est d'ailleurs dans le même ouvrage qu'il propose de mettre sur un pied d'égalité le vivant la machine : « L'organisme vivant est avant tout une machine à vapeur » (Wiener, 2014 [1948], p. 114) ou encore :

Les systèmes nerveux des humains et des animaux, connus comme capables du travail d'un système calculateur, contiennent des éléments idéalement adaptés pour se comporter comme des relais (Wiener, 2014 [1948], p. 225).

La posture cybernétique de Wiener ouvre une brèche dans le dualisme moderne opposant vivant et non-vivant, puis Nature et Société. Nous considérons que cette perspective – rompant avec les dogmes modernes – permet d'intelligibiliser l'Anthropos et la machine sur une base commune. Par exemple l'idée qu'un cerveau fonctionne comme un disque dur ou encore le concept d'intelligence artificielle et d'émotion chez les robots. Du côté du développement scientifique, l'épistémologie cybernétique – ou encore l'épistémologie ou le paradigme de la communication – permet d'intellectuellement joindre deux positions qui semblaient auparavant irréconciliables.

Le but avoué de cette mise en commun – du vivant et non-vivant, en plus d'effacer les frontières entre Nature et Société – dont le méta-langage est le binaire, est de transformer la compréhension de la Nature et de la Société en faveur du concept

d'hyperréalité. L'hyperréel s'inscrit comme la totalité abstraite du monde : « en le réduisant à n'importe quelle opération de mesure que l'on *peut* effectuer [...] en vue d'augmenter l'emprise que l'on a sur [le monde] » (Joly, 2013, p. 33). De surcroît, l'hyperréel est la tentative de simulation d'un monde plus réel que le réel, la simulation en tant qu'abstraction totale :

est [produite] à partir de cellules minituariées, de matrice et de mémoires de modèle de commandement [...] il n'a plus à être rationnel, puisqu'il ne se mesure plus à quelconque instance, idéale ou négative. Il n'est plus qu'opérationnel (Baudrillard, 1981, p. 11).

La mise en place d'une échelle de mesure commune fait en sorte d'aplanir la réalité sur un dénominateur commun, de la rendre calculable et prévisible. Outre l'élément de prévisibilité, la mise en totalité abstraite du monde fait en sorte que la compréhension est remplacée par la subordination de la Nature et du Social sous la forme binaire. Dorénavant régulés par la rationalité de l'efficience, la finalité de la connaissance :

[...] devient la capacité de s'assurer le contrôle de la prévisibilité des conséquences d'une action ou d'une intervention quelconque, pour s'en assurer le contrôle stratégique dans le procès de réalisation de n'importe quel objectif pragmatique (Freitag, 2002, p. 379).

Autrement dit, une fois ordonné sur un dénominateur commun, c'est désormais la subordination du monde naturel et social à toute forme de contingence, mais également du prospectif qui est en jeu. Alors que le projet de contemplation et de compréhension du monde est mis à l'écart, les sciences post-modernes visent à subordonner la réalité en abstraction hyperréel et son prospectif en vue de son contrôle illimité.

2.3 Techno-science : éléments de concrétisation

Les techno-sciences sont constituées d'une série de caractéristiques concrètes et abstraites. Sans chercher à être exhaustifs quant à ces caractéristiques – ce qui reviendrait à effectuer un travail plus que chronophage, voire contradictoire quant aux

différentes théories – nous présentons plusieurs aspects de la complémentarité des techno-sciences.

2.3.1 Les techno-sciences : des équipements de pointe

Une caractéristique nécessaire pour atteindre l'objectif de captation et de subordination de la Nature et du Social au profit de l'hyperréel est la mobilisation d'équipement scientifique à la fine pointe de la technologie. Dans le cadre des techno-sciences, ce sont les délais temporels qui s'effacent au profit de l'instantanéité de la captation et du traitement des données (Chassay, 2006, p. 133). L'accélération dans la captation et le traitement des données offrent un potentiel nouveau en effaçant l'intermédiaire temporel entre le temps de captation et de traitement des données. Autrement dit, l'interstice temporel entre le chercheur et les données captées doit être diminué au maximum pour asseoir le contrôle permanent de l'hyperréel. En prolongement de ce raisonnement, la subordination du réel en hyperréel doit être instantanée, mais doit également dépasser le présent. Il faut désormais avoir la capacité de capter, prévoir et subordonner le futur à partir du présent et du passé.

Les « bons » outils des techno-sciences doivent alors avoir la caractéristique de capter plus, mais surtout de transmettre et de traiter le plus rapidement possible les données captées. Le temps et l'expérience techno-scientifique deviennent alors strictement quantitatifs, au détriment d'un régime d'expérience qualitatif. La techno-science dépouille la conception de la nature de sa contingence en mettant le futur entre les mains de l'Anthropos dans un double mouvement d'abstraction et de quantification du présent et du postérieur en captant l'absolu de façon la plus omnisciente possible.

2.3.2 Les techno-sciences : légitimation stratégique

Bain propose de considérer les techno-sciences en tant processus de légitimation stratégique masquant les alternatives politiques en mettant de l'avant des solutions strictement scientifiques (Bain *et al.*, 2011, p. 160). Autrement dit, si l'entité

gouvernante moderne était l'exécutif, le législatif et le juridique, la gouvernance légitime est dorénavant attribuable à la processualité techno-scientifique. Les décideurs font de plus en plus appel à une légitimité normative processuelle pour établir leur autorité. Cette dernière plonge les aspects politiques et sociaux dans une position latente, au profit d'une problématisation – et bien sûr d'une solution – strictement technique (Bain *et al.*, 2011, p. 162).

Du point de vue discursif, les techno-sciences proposent un discours basé sur une processualité indépendante basé sur des pratiques objectives et précises. Ce discours concernant l'objectivité a pour but de séparer la politique et les questionnements moraux de la scientificité. Les élites politiques et économiques souhaitant maintenir leur position en faisant appel à la Vérité et la Raison relative aux experts et au savoir scientifique pour se distinguer de la populace et de son savoir pratique (Latour, 1999). Alors que les figures de pouvoir mobilisent idéologiquement les techno-sciences pour légitimer leur autorité, la représentation qu'y en est faite est dichotomique :

Those in power use Science (with a capital S) as an ideology, a political weapon against those who may disagree or those who wish to engage in politics and debate (Bain *et al.*, 2011, p. 182).

2.3.3 Les techno-sciences : concrétisation du prospectif fictionnel

Les techno-sciences mettent de l'avant un imaginaire basé sur l'innovation continue. La foi envers l'innovation continue est particulièrement présente du côté de l'informatique et des réseaux de communication. Les techno-sciences incarnent – à travers l'innovation – le pôle de réalisation fictionnelle :

Si la science-fiction a longtemps proposé des récits présentant des technologies utopiques, ces dernières sont quasi systématiquement réalisées par la R&D des grands laboratoires (Michaud, 2017, p. 46).

Autrement dit, la techno-science réalise le prospectif fictionnel. Si au départ, les projets techno-scientifiques peuvent émerger d'une utopie, leur réalisation s'inscrit comme un

passage du fictionnel au réel. La recherche et le développement dans le domaine de l'informatique sont patents de la concrétisation des innovations soutenues dans un secteur à la fine pointe technologique :

These include computer and information sciences as well as all the biosciences and technologies such as molecular biology, genetics, genomics, biotechnology, pharmacogenomics, nanotechnologies, and medical technologies including visualization (Clarke Adele E., 2014, p. 1).

Le savoir est depuis longtemps transformé en application concrète, que ce soit pour améliorer les conditions de vie ou des intérêts commerciaux. Cependant, dans le paradigme techno-scientifique, le savoir fondamental est secondaire :

What is new is the extraordinary expansion of research and the increasingly rapid transformation of knowledge into concrete applications, thereby ensuring continual technological progress. This phenomenon is known as technoscience, and firms play an essential role in it (Woot, 2005, p. 16).

Les techno-sciences proposent de soumettre l'activité scientifique à la captation et la subordination du réel en application concrète préférablement commercialisable. Le support public pour le secteur techno-scientifique est particulièrement présent dans le domaine bio-technologique. Ce champ est considéré comme ayant le potentiel le plus élevé pour effectuer des percées importantes, la recherche et le développement est alors propulsé par une combinaison de fonds public et privé ou par de la recherche effectuée par le public et commercialisé par le privé (Woot, 2005, p. 19). Les secteurs publics et privés justifient cette relation par la possibilité, voire la nécessité, de radicalement transformer la Société et la Nature (Delgado, 2016, p. xi). La technique apparaît alors comme un pôle péremptoire, le support public aux projets techniques se traduit par le support public vers le futur.

Du côté de l'État, le support public à la recherche expérimentale débute dès le 17^e siècle. C'est dans une relation de pouvoir qu'il faut comprendre ce changement. Les sciences modernes sont une condition de possibilité, de création et de découverte de nouvelles entités. Si le contrôle et l'ordonnancement d'entités humaines sont le propre

de l'État traditionnel, son prolongement techno-scientifique est la domestication des entités non-humaines. Sous cet angle il faut : « reconnaître la technoscience pour ce qu'elle est : une formidable source d'amplification du pouvoir » (Séguin, 2012).

CHAPITRE 3

MÉTHODOLOGIE

3.1 Posture du chercheur : le réalisme critique

Dans le cadre de notre recherche, nous adoptons une posture relevant du réalisme critique en science sociale. En ce sens, on postule une relation dialectique entre la structure sociale et l'agentivité du sujet (Price et Martin, 2018, p. 91). Corolairement, en avançant l'idée de structure et d'agentivité – au sens large – nous cherchons dans l'étude de cette dialectique une compréhension coextensive ; comprendre la structure pour comprendre l'agent, comprendre l'agent pour comprendre la structure (Bhaskar *et al.*, 2017).

Notre approche qualitative d'étude de cas s'inscrit dans le réalisme critique dans la mesure où nous cherchons – dans une position herméneutique – à comprendre un phénomène dans un mode réflexif. Le concept de réflexivité est d'ailleurs central dans notre étude. Ce concept est caractérisé par l'attention particulière qu'il faut porter à la résolution des inconsistances – aussi nommées contradictions performatives – entre la théorie et la pratique (le cas à l'étude) (Patel et Pilgrim, 2018). En ce sens, la résolution des inconsistances entre la théorie – ce qui est attendu – et ce qu'on observe – la pratique – est une posture féconde pour concevoir des théories ou en confirmer, etc. On note que cette posture épistémologique est particulièrement intéressante pour étudier des phénomènes ou théories en *construction*.

3.2 Stratégie de recherche : l'étude de cas

Le choix d'une stratégie de recherche – aussi appelé stratégie de vérification – doit être considéré au regard d'un choix général permettant de répondre à une question de recherche. Autrement dit, une stratégie de recherche doit avant tout permettre de confronter une hypothèse à l'épreuve des faits, en faisant ressortir les éléments pertinents pour répondre à un questionnement de recherche (N'Da, 2015, p. 127). On note que le choix d'une stratégie de recherche oriente et limite subséquentement les méthodes de collecte de données.

En l'occurrence, nous favorisons la stratégie de recherche par étude de cas. Fréquemment utilisée dans le domaine des sciences sociales, elle permet de se concentrer sur un objet précis et de généraliser les résultats, voire de confronter des théories (Yin, 2009). L'étude de cas permet la compréhension en profondeur d'un objet en mettant l'accent sur l'analyse d'un phénomène à des fins de généralisations. Elle est à la fois appropriée pour la vérification d'une théorie que pour la construction théorique (Gagnon, 2012, p. 2). En ce qui concerne la position du chercheur, il est en quelque sorte effacé, il n'agit pas sur cas à l'étude, mais mobilise plutôt des méthodes pour comprendre, observer, décrire et interpréter.

L'étude de cas comme stratégie de recherche a trois grandes forces la caractérisant. Elle permet un recueil et une analyse des données en profondeur en tenant compte du contexte dans lequel le phénomène se produit. Comme l'étude peut se concentrer sur un ou plusieurs cas, le développement soutenu d'une généalogie historique du phénomène est favorisé. En dernier lieu, elle assure une validité interne forte, c'est-à-dire que les éléments recueillis disposent d'une aura d'authenticité (Gagnon, 2012).

3.3 La sélection du phénomène et du cas d'étude

Considérant que nous avons choisi d'étudier le phénomène du *Big Data* en science bio-informatique, conjointement à la stratégie par l'étude de cas, nous devons porter notre choix sur un objet significatif – représentatif du phénomène. Cela nous mène à étudier le projet ENCODE, puisqu'il est une figure de proue dans le domaine et qu'il est considéré comme l'archétype d'un projet relatif aux *big sciences*. Il est sans contredit le projet de type *Big Data* en science bio-informatique le mieux structuré, s'inscrivant sur une période significative de plus d'une dizaine d'années et disposant d'une attention importante dans ce même champ d'études (Ecker *et al.*, 2012).

On note que les données du consortium ENCODE ont été citées plus de 6500 fois, ce qui le place dans le 1% des articles les plus cités dans son domaine (SCOPUS, 2019). Ce cas d'étude est donc riche autant au niveau des traces documentaires qu'il laisse, que des effets structurants qu'il a sur le reste du champ de la bio-informatique. En effet, comme nous le verrons ultérieurement, il dispose d'un rôle particulier dans la mesure où il est un producteur de données dont l'objectif est d'agir en tant qu'encyclopédie génétique pour l'ensemble du champ de la bio-informatique.

3.4 La collecte des données : L'étude documentaire

Une collecte des données doit mobiliser des sources variées et en nombre suffisant pour en assurer la validité. Un nombre de documents trop faibles formerait un corpus méthodologiquement questionnable, tandis qu'un nombre trop élevé produirait un effet de saturation des résultats (Yin, 2003). Elle doit également être contextualisée et transparente pour que les résultats puissent être reproduits et analysés.

Dans notre cas, nous favorisons la collecte de données documentaires. Le terme « document » fait référence aux sources de renseignements préexistants accessibles au chercheur (N'Da, 2015, p. 129). Les documents peuvent être disponibles sous plusieurs formes, dans le cas de cette recherche, ce sont les documents écrits qui nous intéressent. La documentation sélectionnée en vue de cette recherche a la caractéristique d'avoir été produite par le consortium ENCODE. En ce sens, les sources documentaires n'ont pas été élaborées en vue d'une recherche, elles constituent plutôt des traces de l'activité socio-scientifique.

Nous procédons à l'analyse de divers documents relatifs au projet ENCODE, comme les devis de subventions, les politiques de dissémination & d'utilisation des données, puis l'article scientifique de référence du consortium. Ces documents permettent une couverture fiable sur une période temporelle donnée et contiennent des éléments précis.

Tableau 3-1 Composition du corpus

Sources	Types	Informations	Période	Nombre de documents
Nature / Consortium ENCODE	Article scientifique	Organisation, orientations des recherches, objectifs	Publication en 2012	1
Consortium ENCODE	Politique d'utilisation et de distribution des données	Rôle et accessibilité des données	Publication en 2014	2
U.S. Department of Health & Human Services (NIH)	Funding Opportunity Announcement	Financement, organisation, projet en cours, objectifs	Émission des documents en 2016	5

3.5 Traitement et analyse des données

Trois étapes itératives constituent habituellement le traitement des données ; l'épuration, le codage et l'analyse. Bien que présentées sous un format séquentiel, il est nécessaire de revenir plusieurs fois à chacune d'entre elles pour produire un cadre d'analyse intellectuellement et méthodologiquement achevé (Gagnon, 2012, p. 72).

Dans notre cas, le processus d'épuration a été effectué de manière à filtrer les sections des documents n'ayant que peu d'intérêt dans le cadre de notre projet de recherche. Par exemple, des sections étaient communes à chacun des appels à candidatures pour obtenir du financement dans le cadre du projet ENCODE. Les sections des documents n'étant pas pertinentes dans le cadre de notre analyse ont également été exclues.

Pour faire ressortir les éléments pertinents à notre analyse – donc pour former notre corpus – nous avons effectué des synthèses documentaires. Procéder sous la forme de synthèse nous permet entre autres d'être souple quant aux éléments que nous avons sélectionnés, tout en nous permettant d'y inscrire des éléments nécessitant d'être contextualisés ou expliqués.

Quant à l'analyse de notre corpus, elle est effectuée en interprétant les synthèses documentaires à partir des catégories représentant la post-modernité scientifique. En ce sens, le *tableau 3.2 grille thématique* permet de dresser un portrait des thématiques abordé lors de l'analyse du corpus.

Tableau 3-2 Grille thématique

Caractéristiques	Science post-moderne ou techno-science
<i>Compréhension</i>	Corrélation ; Quantitatif ;
<i>Méthode</i>	Inductive ; Abductive ; Déductive ;
<i>Objectif / finalité</i>	Prédiction ; Appliqué ; Efficacité ; Subordination du réel présent et prospectif au profit de la notion d'hyperréel ;
<i>État du savoir</i>	Traduction communicationnelle en langage informatique ou quantifiable ;
<i>Réalisation</i>	Du fictionnel vers le réel ;
<i>Ambition</i>	Les sciences au service du marché pour la création de la valeur ;
<i>Contexte de reproduction sociale</i>	Décisionnel-opérationnel ;

Le présent tableau dresse, *grosso modo*, les catégories représentatives de la post-modernité, telle que définie dans notre cadre théorique. Comme la post-modernité est un concept à la fois non-fini et corolairement polysémique, les catégories extraites de notre cadre théorique permettent de mettre l'accent sur des points d'intérêt qui pourront être revisités au cours de la discussion.

3.6 Limites et réflexions

Par souci de transparence intellectuelle et de reconnaissance des limites de notre étude de cas, nous proposons de traiter brièvement de celle-ci en faisant part de quelques-unes de nos réflexions quant à notre recherche. Il est impératif de prendre en compte que l'étude de cas est la représentation ponctuelle d'une réalité médiée, la généralisation des résultats de l'étude de cas doit se faire en tenant compte de la

dialectique sujet/théorie. En étudiant un phénomène en profondeur, les données et l'analyse font ressortir des éléments spécifiques qu'il faut analyser en tenant compte d'un contexte plus large. Or, en formant une théorie générale à partir d'un objet spécifique, l'extrapolation de ces résultats à d'autres cas doit se faire avec prudence.

Comme nous l'avons précisé à la section *3.1 Posture du chercheur : le réalisme critique*, la dialectique structure/agentivité et théorie/pratique donne un potentiel analytique important tout en accentuant les complexités. Si la théorie s'éloigne du réel médié, faudra-t-il la critiquer en proposant des modifications, la rejeter en bloc ou créer un cadre théorique plus représentatif de la réalité ? Si la pratique n'est pas représentative de la théorie, est-ce simplement un écart herméneutique, une faiblesse méthodologique ou le témoignage d'une étude poussant la théorie à ses limites?

CHAPITRE 4

PRÉSENTATION DES RÉSULTATS

Ce chapitre porte sur la présentation des résultats de la cueillette de données. Pour procéder à l'étude du projet ENCODE, nous avons recueilli trois types de documents en décrivant leurs tenants et aboutissants. La présentation des documents se fait sous la forme de résumé-synthèse dont l'interprétation théorique est différée au chapitre suivant.

Premièrement nous avons recueilli les *Funding Opportunity Announcement* (FOA) annoncés par le *National Institutes of Health* (NIH) et sa sous-branche le *National Human Genome Research Institute* (NHGRI). Ces annonces sont riches en informations, car elles présentent les intentions du consortium pour les années à venir. Ces documents contiennent entre autres une feuille de route générale pour le projet ENCODE, mais également des attentes spécifiques à chacun des pans du projet, le financement prévu et le mode d'organisation.

Deuxièmement, le consortium a produit de la documentation concernant l'utilisation et la publication de ses données. Les *Data Release Policy & Data Use Policy* offrent la possibilité de comprendre le rôle des données bio-informatiques pour le consortium et sur la diffusion des connaissances.

En dernier lieu, l'article scientifique *An Integrated Encyclopedia Of DNA Elements In The Human Genome* publié en 2012 dans *Nature* est analysé dans le cadre de cette

recherche. En tant qu'article de référence du projet, il couvre les réalisations passées et les buts du projet ENCODE.

4.1 Funding Opportunity Announcement

Coordonnés par le *National Institutes of Health* et le *National Human Genome Research Institute*, les *Funding Opportunity Announcement* présentent les possibilités de financement pour les groupes intéressés à se joindre au projet. L'annonce des FOA s'effectue en bloc avant que la phase précédente soit achevée, permettant un temps minimum de transition entre les phases du projet. À la suite de l'appel initial, les groupes intéressés envoient une candidature pour intégrer l'équipe du consortium.

Nous présentons les FOA pour la quatrième phase du projet ENCODE, s'étalant de 2017 à 2020. Les *Funding Opportunity Announcement* de cette phase ont été diffusés en janvier 2016, les premières subventions débutant au plus tôt en décembre 2016, et ce pour une période de 4 ans (National Institutes of Health, 2016c). La quatrième phase du projet est la plus récente et la plus achevée, elle permet de tenir compte des recherches précédentes et des intentions du consortium. Dans le cadre de cette phase d'expansion du projet, 5 *Funding Opportunity Announcement* sont diffusés.

Le consortium ENCODE nécessite que des groupes aux compétences diversifiées se partagent des tâches complexes, et ce dans une logique d'ensemble. Chaque pan du consortium doit donc répondre à des critères précis, puisque les tâches sont divisées selon les compétences des groupes participants et les nécessités du projet. Selon les FOA diffusés pour la phase 4 du projet, le travail des membres du consortium sera divisé en 5 pans constitués de 1 à 10 unités et dont le budget total oscille entre 30.9\$ millions à 38\$ millions par années pour une période de 4 ans.

Tableau 4-1 Présentation des *Funding Opportunity Announcement*

<i>Funding Opportunity Announcement</i>	Résumé	Subventions et projets
FOA RFA-HG-16-002: Expanding the Encyclopedia of DNA Elements (ENCODE) in the Human and Mouse	Production à grande échelle de données génétique à partir d'échantillons biologiques	\$15.5-20 M ; 6 à 8 unités ;
FOA RFA-HG-16-003: Characterizing the Functional Elements in the Encyclopedia of DNA Elements (ENCODE) Catalog	Développer et expérimenter des méthodes pour étudier les fonctions génétiques	\$5.9-6.5 M ; 7 à 10 unités ;
FOA RFA-HG-16-004: Computational Analysis of the Encyclopedia of DNA Elements (ENCODE) Data	Développement et application de méthode informatique pour analyser les données Analyse des données et recommandations sur les données ayant un potentiel analytique	\$2.5-3 M ; 4 à 6 unités ;
FOA RFA-HG-16-005: ENCODE Data Coordinating Center	Coordination générale Hébergement et distribution des données	\$5-5.5 M ; 1 unité ;
FOA RFA-HG-16-006: ENCODE Data Analysis Center	Coordination analytique	\$2-3 M ; 1 unité ;

4.1.1 FOA RFA-HG-16-002: Expanding the Encyclopedia of DNA Elements (ENCODE) in the Human and Mouse

Cette section du projet porte sur le *mapping* des données à partir d'échantillons biologiques. Les laboratoires obtenant cette subvention ont le mandat de produire des données bio-informatiques en quantités importantes, de façon peu onéreuse et dont la fiabilité est de haut niveau :

The primary activity of the Mapping Centers will be data generation in an expanded number of cell contexts (cell types, fates, states) relative to those studied to date in ENCODE. For information on existing ENCODE data and biological samples assayed (National Institutes of Health, 2016a, p. 6).

Les données produites pour enrichir la base de données encyclopédique doivent provenir majoritairement du génome humain et minoritairement de celui de souris. On note que les équipes de *mapping* sont uniquement des producteurs de données, mais n'ont pas de mandat analytique.

Les équipes de production de données sont encouragées à obtenir un nombre suffisant d'échantillons pour les partager avec les autres laboratoires à des fins de vérification. Elles doivent se coordonner avec les autres groupes pour ne pas répéter les mêmes expérimentations. Les méthodes et résultats de production doivent s'harmoniser avec les précédents à des fins de reproductibilité des expérimentations et pour créer une encyclopédie cohérente (National Institutes of Health, 2016a).

4.1.2 FOA RFA-HG-16-003: Characterizing the Functional Elements in the Encyclopedia of DNA Elements (ENCODE) Catalog

Potentiellement le pan du projet impliquant le plus d'équipes – 7 à 10 selon les candidatures – cette section d'ENCODE a pour objectif d'identifier des fonctions génétiques au sein de bio-contextes variés, par exemple des maladies ou des anomalies génétiques. Comme les recherches du consortium ont élargi la notion de fonction

généétique au-delà des protéines encodantes, de nouvelles méthodes pour comprendre les réactions entre des éléments génétiques doivent être développées :

The focus will be to develop and test strategies that could be broadly applied by the community, and to begin to ask how to make better predictions of function, rather than to attempt to characterize the entire catalog of candidate functional elements in all cell contexts (National Institutes of Health, 2016b, p. 6)

Axées principalement sur le développement de procédures de recherche biologique, les équipes devront développer des modes d'expérimentation pour prévoir les fonctions des éléments biologiques étudiés. Limité aux données biologiques présentes dans l'encyclopédie ENCODE, il s'agit d'étudier les éléments pour offrir de nouvelles méthodes de généralisation des procédés expérimentaux, ceux-ci étant pour la plupart peu développés concernant les fonctions non-encodantes (National Institutes of Health, 2016b).

4.1.3 FOA RFA-HG-16-004: Computational Analysis of the Encyclopedia of DNA Elements (ENCODE) Data (U01)

Les 4 à 6 groupes de cette équipe ont pour mandat de développer et appliquer des méthodes d'analyse informatique à partir des données de l'encyclopédie génétique du projet ENCODE. L'objectif est d'améliorer les capacités d'interprétation des données existantes en cherchant par exemple des corrélations entre les éléments de l'encyclopédie ou en combinant celles-ci avec des données externes :

A proposed research project should, as necessary, include the development of software with thorough documentation that would allow a broad cross-section of researchers to readily use that software. For the same reason, software developed by awardees should be modular, extensible and robust, so that it can work as a stand-alone package and be adopted and further developed by the community (National Institutes of Health, 2016c, p. 7).

Ces pistes corrélatives ont pour but de dresser une feuille de route sur les cellules à analyser en plus d'étudier le lien entre certaines maladies et des cellules de l'encyclopédie. On retient que le mandat de ce pan du projet s'axe principalement sur

l'informatique, la recherche de corrélation et la production de logiciel et de formules d'analyse (National Institutes of Health, 2016c).

4.1.4 FOA RFA-HG-16-005: ENCODE Data Coordinating Center (U24)

Le *Data Coordinating Center* (DCC) du projet ENCODE gère la distribution et l'organisation des données génétiques. En tant que ressource principale de la communauté intéressée par les données de l'encyclopédie, le DCC s'assure qu'un portail et des outils sont disponibles pour que des recherches externes profitent de son travail. Cela fait en sorte qu'il doit maintenir la base de données précédente et s'assurer qu'elle est compatible avec les futures données du consortium. Au-delà de ses tâches de coordination générale du consortium, il s'assure de contrôler la qualité du travail des membres, uniformise les données, puis les publie le plus rapidement possible :

- manage the uptake, quality control, and accessioning of data, metadata, tools, and analyses from all members of the ENCODE Consortium, as well as relevant functional genomics data generated outside of the consortium
- provide a portal to allow the community to visualize and download data, tools and analyses
- serve as a coordinating center for consortium activities
- lead consortium outreach efforts (National Institutes of Health, 2016d).

4.1.5 FOA RFA-HG-16-006: ENCODE Data Analysis Center (U24)

En tant que pôle de coordination analytique du consortium, le *Data Analysis Center* (DAC) ne produit pas les données, mais les interprètes et les classes dans l'encyclopédie. Il effectue un suivi et guide le développement des outils et logiciels nécessaires au fonctionnement du catalogue ENCODE. Comme les données de l'encyclopédie proviennent d'un nombre important de laboratoires, et ce depuis 2001, le DAC produit des logiciels de standardisation :

The DAC will be a member of the AWG, and will work with the informatics components of the individual ENCODE data production centers and AWG to coordinate and facilitate integrative analyses of the ENCODE and related data.

The DAC will provide resources and coordination activities that will enable the ENCODE Consortium to provide useful information about functional sequence elements in the human and mouse genomes to the larger biomedical research community (National Institutes of Health, 2016e, p. 7)

Son travail analytique doit servir de base pour guider les laboratoires de génération des données en identifiant les cellules et les données ayant une « valeur » pour faire de nouvelles découvertes à court et long terme. Par exemple, les données permettant de mieux comprendre des corrélations relatives à des maladies (National Institutes of Health, 2016e).

4.2 Politique d'utilisation & de distribution des données

La seconde section porte sur la politique d'utilisation & distribution des données de l'encyclopédie génétique. Cette documentation permet d'étudier le rôle des données dans le processus scientifique d'ENCODE. Basée sur une politique des données ouvertes, l'Encyclopédie génétique favorise la dissémination des données pour une utilisation par des projets externes. L'accessibilité aux données n'est soumise à aucune barrière financière et leur publication dans des bases de données publiques s'effectue sans moratoire quant à leur utilisation. Également, il n'y a aucune restriction quant à la quantité de données téléchargeable et c'est l'ensemble des données qui sont disponibles, et ce, gratuitement. En mettant de l'avant des principes de sciences ouvertes¹⁰, le consortium souhaite que des groupes externes bénéficient des données pour mener leurs propres recherches (ENCODE Project Consortium, 2014).

¹⁰ Le principe de science ouverte tel que développé par Robert K. Merton, au cours des années 1960-70, met de l'avant une forme de communisme scientifique où le savoir est considéré comme un commun. Dès lors, l'éthos scientifique devrait mettre de l'avant un régime où les découvertes et le processus scientifiques représentent un idéal politique de collaboration social incompatible avec le capitalisme et la propriété privée (Merton, 1968). Or, le terme peut également être considéré comme un artéfact du régime néolibéral de génération des données se conformant aux impératifs de marché ou encore par rapport à un régime démocratique atrophie (Mirowski, 2018). Alors que Merton soulignait l'incompatibilité de l'accaparement capitaliste, l'*open science* est de nos jours – plus souvent qu'autrement – synonyme d'un régime de production des données lié à un processus de gouvernamentalité algorithmique. Autrement dit, l'*open science* est un terme polysémique faisant d'un

S’inscrivant dans une logique de continuité avec le principe de science ouverte du projet, la politique d’utilisation des données est composée d’une seule page aisément compréhensible. Elle indique qu’il est impératif de citer les publications *An Integrated Encyclopedia Of DNA Elements In The Human Genome* (2012) et le numéro de référence du *dataset* utilisé (ENCODE Project Consortium, 2019a). Il est ainsi possible de retracer les publications ayant recours aux données de l’encyclopédie génétique. Également, cette politique est très simple pour servir son objectif qui est de favoriser l’utilisation des données par des groupes externes.

Tableau 4-2 Synthèse thématique de la politique d’utilisation et de distribution des données

Thématique	Synthèse
Rôle des données	Production pour utilisation par des acteurs externes au projet.
Accessibilité des données	Données ouvertes et accessibles gratuitement

4.3 Article scientifique de référence : An Integrated Encyclopedia Of DNA Elements In The Human Genome

L’article *An Integrated Encyclopedia Of DNA Elements In The Human Genome*, publié en 2012 par les membres du consortium ENCODE, est la publication de référence du projet. À ce titre, il est nécessaire de le citer pour utiliser les données de l’encyclopédie génétique. Ce moyen de faire permet de retracer aisément les articles ayant utilisé les données de l’encyclopédie génétique. On note que depuis la parution de l’article en 2012, celui-ci a été cité près de 6500 fois selon la base de données SCOPUS, ce qui le place dans le 99^e percentile pour le nombre de citations dans le champ médical

côté référence à un idéal démocratique où le processus scientifique est considéré comme un commun en opposition à la propriété privée. De l’autre, il peut faire référence à un régime de génération de données où plus il y aurait de données rendues publiques, plus il y aurait de *démocratie*.

(SCOPUS, 2019). En date d'octobre 2015, l'encyclopédie génétique publiée par le consortium compte environ 3500 expérimentations examinant plus de 300 types de cellules humaines et plus de 1000 expérimentations sur 150 types de cellules de souris (National Institutes of Health, 2016a).

À long terme, l'objectif d'ENCODE est d'identifier et de séquencer l'ensemble des éléments génétiques humains ayant des fonctions, en plus de les mettre à disposition de la communauté scientifique. Étant donné que peu de ressources sont déployées dans le monde scientifique pour annoter des régions génomiques non associées à des protéines, le projet est d'autant plus important, puisqu'il comble un vide scientifique en ce domaine.

D'entrée de jeu, l'article précise que le génome humain : « encodes the blueprint of life », c'est-à-dire les plans pour ne créer rien de moins que la vie (ENCODE Project Consortium, 2012, p. 57). Le génome humain est donc considéré comme une séquence qui : « [...] provides the underlying code for human biology » (ENCODE Project Consortium, 2012, p. 57). Envisagé comme tel, le génome humain est alors le code à la base de la vie. Cela mène le consortium à vouloir encoder l'ensemble des éléments ayant des effets les uns sur les autres : « The Encyclopedia of DNA Elements (ENCODE) project aims to delineate all functional elements encoded in the human genome » (ENCODE Project Consortium, 2012, p. 1).

On mentionne que l'accomplissement de ce but est rendu scientifiquement et techniquement possible parce que : « The advent of more powerful DNA sequencing technologies now enables whole-genome and more precise analyses with a broad repertoire of functional assays » (ENCODE Project Consortium, 2012, p. 1).

Ceci est d'ailleurs rappelé en conclusion de l'article :

Importantly, for the first time we have sufficient statistical power to assess the impact of negative selection on primate-specific elements, and all ENCODE classes display evidence of negative selection in these unique-to-primate elements (ENCODE Project Consortium, 2012, p. 71)

Dès lors, le consortium reconnaît que la technique – dans ce cas le *Big Data* – est une de ses conditions de possibilité. En ce qui a trait à la valeur scientifique de l'encyclopédie génétique, le consortium souligne l'apport des données générées :

The unprecedented number of functional elements identified in this study provides a valuable resource to the scientific community as well as significantly enhances our understanding of the human genome (ENCODE Project Consortium, 2012, p. 71).

Autrement dit, les données générées le sont principalement pour servir de ressources à d'autres projets, en quelque sorte comme base de données pour des projets externes. Les données générées deviennent pertinentes dans la mesure où elles sont utilisées par des groupes ou chercheurs externes au projet. Pour arriver à fournir une encyclopédie génétique favorisant les corrélations et l'analyse entre les données, le consortium précise :

For consistency, data were generated and processed using standardized guidelines, and for some assays, new quality-control measures were designed (ENCODE Project Consortium, 2012, p. 71).

En ce sens, des procédures et manières de standardiser le travail des groupes sont nécessaires pour que l'encyclopédie génétique offre des données homogènes, donc comparables entre elles.

Tableau 4-3 Synthèse thématique de l'article de référence du projet ENCODE

Thématique	Synthèse
Objectif	Identifier et séquencer l'ensemble des éléments du génome humain ayant des fonctions
Épistémologie	La vie sous forme de code génétique informatisable
Méthode	Corrélation pour déterminer les éléments à produire et standardisation des méthodes de travail
Condition de possibilité	<i>Big Data</i>
Données	Les données en tant que ressources scientifiques

CHAPITRE 5

DISCUSSION : DE LA *TECHNÈ* À L'*ÉPISTÉMÈ*, DE L'*ÉPISTÉMÈ* À LA *DOXA*

Nous avons problématisé notre recherche en expliquant les tenants et aboutissants du *Big Data* selon l'angle technique et scientifique. En ce qui concerne l'évolution de celui-ci dans le domaine des sciences, notre choix s'est arrêté sur le champ des sciences biologiques, puisque celui-ci s'est rapidement transformé en domaine bio-informatique¹¹. Étant donné que le projet ENCODE s'inscrit comme figure proue dans ce domaine et que la littérature scientifique s'intéresse à celui-ci, il a naturellement été sélectionné comme objet de recherche.

Cette recherche est contextualisée à partir d'un cadre théorique permettant d'intelligibiliser – au vu de la post-modernité scientifique dont la matrice technoscientifique est en filiation avec la cybernétique – les sciences et la technique. La mobilisation d'un cadre théorique de nature sociologisante présente des avantages et des inconvénients. Dans notre cas, le principal avantage est de pouvoir mettre en perspective les changements opérés par la technique sur le processus scientifique, en tenant compte de repère socio-historique précédent, ouvrant la porte à une généralisation de nos résultats. Le principal inconvénient de cette approche est la possibilité de perdre de vue notre objet de recherche. Pour pallier à cet inconvénient, tout en favorisant une compréhension plus large du phénomène, nous divisons notre discussion en deux parties successives.

¹¹ On note que bien que le domaine bio-informatique prenne une place considérable dans le domaine de la biologie, il ne le constitue pas entièrement. En effet, le domaine est constitué de plusieurs champs de recherche où plusieurs méthodes et épistémologies coexistent.

La première section s'intitule : *De la technè à l'épistémè: Big Data et science*. Elle porte sur une interprétation des données de notre recherche en posant notre regard sur les implications du *Big Data*. Cette première section est divisée en cinq thématiques : la nature du savoir, l'état du sujet, le mode de compréhension, les ambitions et conditions de possibilité technique et les modalités organisationnelles.

La seconde section *De l'épistémè à la doxa : cybernétique et société* traite de la transition de l'épistémologie scientifique à la *doxa*. Comme les sciences sont le principal moteur de compréhension du monde et que cela se répercute activement et passivement dans la société, étudier l'épistémologie scientifique cybernéticienne ouvre la porte à une compréhension de la *doxa*. On y aborde principalement la thématique de la réification scientifique sous le paradigme cybernéticien pour comprendre la *doxa* post-moderne.

5.1 De la *technè* à l'*épistémè* : *Big Data* et science

Dans le cadre de cette section de la discussion, on s'intéresse aux effets des *Big Data* sur les sciences et plus précisément dans le cadre du projet ENCODE. Pour introduire les 5 thématiques de cette section, nous abordons initialement les concepts de *technè* et d'*épistémè*.

Le concept de *technè*¹² – dont les racines remontent à la Grèce antique – fait référence à une manière de faire, une méthode appliquée ou encore à l'artisanat (Liddell et Scott, 1940). Platon ne faisait pas nécessairement référence à des artefacts techniques quand il utilisait le terme *technè*, le terme s'appliquait autant aux techniques d'agriculture qu'à l'art de la politique (Winner, 2002). La *technè* se distingue notamment de

¹² τέχνη en grec antique.

l'*épistémè*¹³ – concept popularisé par Foucault dans *Les mots et les choses: une archéologie des sciences humaines* (1966) – faisant entre autres référence aux formes de savoirs et de compréhension. Alors qu'au cours de la période antique, *épistémè* et *technè* n'étaient pas en relation dialectique dont le dessein est la production capitaliste, le développement scientifique est dorénavant enchâssé avec celui de la technique.

Sans tomber dans le piège du déterminisme où la technique serait l'élément déterminant de l'organisation sociale, on constate que le *Big Data* est une condition de possibilité et influence plusieurs recherches scientifiques, dont le projet ENCODE. En ce sens, étudier le développement du *Big Data* permet de rendre compte d'une technique ayant une influence particulière sur le développement scientifique. L'utilisation du *Big Data* au cours d'une recherche scientifique joue un rôle structurant dans le processus de recherche, bien qu'il puisse varier d'un projet à l'autre.

5.1.1 De l'encyclopédie à la base de données : vers la mise à mort de l'encyclopaedia?

Données et connaissances ne sont pas synonymes, les données – qu'elles soient en quantité limitée ou virtuellement illimitée – n'ont pas de valeur intrinsèque et ne sont pas proportionnelles à notre compréhension universelle du monde social et naturel. C'est intégré à un système, qu'il soit économique, politique ou scientifique, qu'elles acquièrent un sens ou une valeur. En tant que condition de possibilité, le *Big Data* est un moyen de produire et d'échanger – dans une mesure exponentielle – des données, mais ne multiplie pas nécessairement le savoir ou la compréhension d'un phénomène.

Alors que l'extrait principal du projet ENCODE n'est pas des savoirs ou des connaissances, mais bien des données, il importe de remettre en contexte leur rôle dans une perspective scientifique. Conjugué à une politique de données ouvertes limpide, l'objectif d'ENCODE est de fournir des données qui seront utilisées par d'autres

¹³ ἐπιστήμη en grec antique.

acteurs du milieu scientifique ou encore des corporations ayant l'intention de breveter le vivant. Certes utiles pour plusieurs groupes, les données génétiques d'ENCODE ne constituent pas à proprement parler des connaissances. Pour comprendre cette posture, il est pertinent de s'intéresser à la définition générale d'une encyclopédie :

Ouvrage où l'on expose méthodiquement ou alphabétiquement l'ensemble des connaissances universelles (encyclopédie générale) ou spécifiques d'un domaine du savoir (encyclopédie spécialisée). Dans un dictionnaire encyclopédique, développement qui présente une information structurée sur le concept défini en entrée (Larousse, 2019).

Contrairement à une base de données, une encyclopédie est un recueil de connaissances constitué de concepts et de théories en rapport avec un sujet ou un domaine d'étude. Bien que présenté comme tel, *l'Encyclopedia of DNA Elements* ne convient pas à proprement parler aux critères d'une encyclopédie. Pour considérer ENCODE de cette façon, il faudrait appréhender les sciences génétiques quasi exclusivement sous l'angle des données, puis les confondre avec les connaissances. Or, il apparaît que pour le consortium, ses données ont une valeur scientifique de compréhension du corps sous l'angle bio-informatique. Cependant, elles demeurent des données nécessitant d'être interprétées à partir de concepts.

Cela fait écho à la post-modernité scientifique où la forme de savoir valide doit nécessairement s'inscrire dans le langage binaire. C'est ainsi que la prédiction de Lyotard semble se réaliser :

Dans cette transformation générale, la nature du savoir ne reste pas intacte. Il ne peut passer dans les nouveaux canaux, et devenir opérationnel, que si la connaissance peut être traduite en quantité d'information. On peut donc en tirer la prévision que tout ce qui dans le savoir constitué n'est pas ainsi traduisible sera délaissé, et que l'orientation des recherches nouvelles se subordonnera à la condition de traduisibilité des résultats éventuels en langage de machine. (Lyotard, 1979, p. 13).

En effet, la nature de la connaissance et du savoir est transmuée jusque dans son essence. Alors que l'*épistémè* moderne considérait encore la connaissance sous l'angle

de concepts, théories et explications, un glissement s'est opéré en catimini. Dans la post-modernité, les encyclopédies sont remplacées par les bases de données en langage binaire. Cette mutation modifie les principes de la connaissance du monde social et naturel, où la quantité d'information disponible semble proportionnelle à notre compréhension de ce dualisme. Ainsi, si certaines encyclopédies avaient la prétention de détenir l'ensemble des connaissances d'une discipline, on peut avancer que les bases de données les plus volumineuses, variées et dont les données sont plus rares, auront des aspirations similaires.

De surcroît, Nature et Social sont remplacés par l'hyperréel, qui on le rappelle, est une forme de simulacre basé sur le langage binaire ayant la prétention d'être plus vrai que le réel lui-même. Le langage binaire permet d'intelligibiliser l'entièreté du monde selon un dénominateur commun. Dès lors, il advient qu'un procès de production de données est nécessaire à la compréhension et la formation de l'hyperréel. Dans le cas de notre objet de recherche, la biologie se transforme en science bio-informatique et le corps humain et animal est réifié sous le paradigme cybernétique.

5.1.2 *État du savoir : Le sujet informationnel*

L'expression *biologie moléculaire* doit son origine au mathématicien et physicien Warren Weaver au cours des années 1930. C'est avant tout une volonté de transposer à la fois la méthodologie et la rigueur théorique relative aux sciences de la nature qui anime le scientifique. La naissance de ce champ de recherche est suivie au cours des années 1950 – donc dans la période de l'après-guerre – par un intérêt grandissant pour la physique nucléaire et la biologie moléculaire (Lafontaine, 2004). Cet élément se décline en une ontologie où le sujet est conçu sous une forme informationnelle dans le domaine des sciences de la vie. La découverte de l'ADN dans les années 1950 prolonge l'idée d'un code bio-chimique, à l'instar d'un code informatique, qui programme l'individu. Au cours des années 1990, le *Human Genome Project* fondait plusieurs espoirs pour enfin percer le *code* génétique humain. Le prix Nobel de chimie, Walter

Gilbert, déclarait que ce projet allait enfin permettre de transporter un être humain sur un CD (Kay, 2000). Cette brève introduction nous permet de contextualiser les rapides changements ayant constitué la discipline des sciences bio-moléculaires, puis sa transition en science bio-informatique.

Les sciences de la vie ont donc bel et bien emprunté le chemin d'une rigueur théorique et méthodologique relative aux sciences de la nature. L'étude du projet scientifique ENCODE démontre que celui-ci considère que les séquences génétiques sont à la base de la vie : « The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown » (ENCODE Project Consortium, 2012, p. 57). La vie est alors conçue de façon informationnelle, les structures bio-chimiques étant désormais le code à la base de la vie humaine et animale.

Si cette affirmation va de soi dans le domaine de la bio-informatique, ces implications – au regard du *Big Data* – ont techniquement progressé sans que des questionnements épistémologiques critiques fassent contrepoids. Le développement scientifique et technique s'en tient habituellement à l'examen des conséquences à court terme sans se soucier des implications à long terme. Autrement dit, si une réalisation est techniquement possible, les chances qu'elle ne se concrétise pas sont minces, particulièrement à la vue de considérations épistémologiques.

Or, le paradigme du sujet informationnel a des incidences ontologiques et scientifiques majeures. Les sciences de la vie se convertissent en sciences informatiques – de là l'appellation bio-informatique – dont l'objectif est à terme une compréhension totale de l'Être sous forme binaire, brouillant du même coup les frontières entre le vivant et le non-vivant. La forme de savoir qui devient exclusivement épistémologiquement valide est alors quantitative. Sous cet angle, il est impératif de procéder à la conversion du vivant et non-vivant sous le dénominateur commun du langage binaire. Autrement dit, on assiste à un processus de choséification – sous le paradigme cybernéticien – du corps humain, voire de la vie de façon généralisée.

Bien que le procès de réification fasse partie du prélude bio-moléculaire, le *Big Data* permet de se rapprocher techniquement de l'informatisation totale du corps. On passe de l'échantillon génétique isolé à l'aspiration de compréhension complète du corps humain sous langage binaire. Ainsi considérée, l'activité scientifique s'intéressant au corps humain nécessite l'utilisation d'une puissance informatique toujours plus performante et permet à terme d'arriver à concevoir une compréhension unitaire de celui-ci. Le rôle de la technique sous la forme du *Big Data* devient alors primordial pour le développement des *sciences de la vie*.

5.1.3 Forme de compréhension et méthode : La corrélation en tant que moyen et fin

La raison d'être du consortium ENCODE est la production de données bio-informatiques pour définir tous les éléments fonctionnels encodés dans le génome humain. En tant que producteur de donnée, le consortium dispose donc d'un rôle analytique restreint, axant plutôt ses objectifs de recherche sur la création et l'organisation de données bio-informatiques. Le choix des données génétiques à produire est déterminé selon les corrélations entre les différentes données, celles offrant des pistes prometteuses pour faire des liens entre différentes variables ayant des fonctions génétiques sont sélectionnées : « ENCODE has identified different classes of candidate functional regions based on their correlation with biochemical signatures, but has not tested the majority of those to determine what, if any, biological roles they actually play » (National Institutes of Health, 2016b, p. 2).

Or, si la corrélation peut servir d'étincelle pour des recherches dont l'objectif est la détermination d'une théorie générale ou d'une explication causale, le cas d'ENCODE diffère de ce modèle. En effet, comme la production de données bio-informatiques est l'intérêt principal du projet, le modèle corrélatif devient un moyen et une fin en soi. Puisque les capacités de génération et de conversion des données biologiques en code binaire sont techniquement et économiquement limitées, le modèle corrélatif permet

de guider le projet. Autrement dit, le mode de recherche corrélatif permet d’orienter les futures recherches en identifiant des portions de données biologiques pertinentes à transformer sous forme de données bio-informatiques.

5.1.4 Ambitions & conditions de possibilités

Les sciences et la technique – nous tenons à le rappeler – se fondent mutuellement. Les limites imposées par les possibilités techniques s’incarnent réciproquement dans les structures mentales des prospectives scientifiques. Ainsi le développement scientifique s’effectue à l’intérieur de limitation technique et fait du même coup progresser ces limites selon une logique de boucle.

Tel que mentionné dans l’article de référence du projet ENCODE, les réalisations scientifiques de celui-ci sont rendues possibles par les avancées techniques dans le domaine de l’informatique et l’électronique :

The advent of more powerful DNA sequencing technologies now enables whole-genome and more precise analyses with a broad repertoire of functional assays (ENCODE Project Consortium, 2012, p. 57).

Importantly, for the first time we have sufficient statistical power to assess the impact of negative selection on primate-specific elements, and all ENCODE classes display evidence of negative selection in these unique-to-primate elements (ENCODE Project Consortium, 2012, p. 71).

Ces références à la puissance statistique et informatique sont reliées au développement technique des *Big Data* en science bio-informatique. Si autrefois la production, le stockage et le partage d’un imposant corpus de données étaient complexes, les laboratoires d’ENCODE peuvent désormais produire rapidement des données informatiques à partir d’échantillons biologiques. D’ailleurs, les équipements produisant les données génétiques informatisées occupent maintenant moins d’espace physique que les centres de données les hébergeant.

Comparativement à des projets scientifiques plus modestes, on constate que le volume de données est significatif, voir qu'il peut être considéré comme proportionnel au succès du projet. En effet, le projet a en date de 2017, produit plus de 500 terabytes de données bio-informatiques portant sur le code génétique humain et animal (Davis *et al.*, 2018). En référence aux principes techniques du *Big Data* – le volume, la vitesse et la variété – on constate que le volume est le principal aspect technique qui caractérise le projet. Nous soulignons que les coûts de production des données sont relativement raisonnables, étant de 15 à 20 millions \$ par années pour 6 à 8 unités de recherche. Le *Big Data* est donc une condition de possibilité technique rendant possible les aspirations d'ENCODE à atteindre un volume de donnée intégral quant au génome humain.

5.1.5 Du groupe au consortium, du chercheur au gestionnaire ;

Le mode organisationnel du projet ENCODE est éclairant à plusieurs égards pour comprendre les implications du *Big Data* dans le domaine scientifique. Avant de l'étudier, il importe de préciser que pour cette section, on doit mettre en contexte les buts du projet et les limitations techno-scientifiques de celui-ci. En ce sens, nous analysons l'organisation du projet ENCODE à la vue des éléments qui nous apparaissent comme déterminant pour son organisation.

Dans le cas d'ENCODE, nous rappelons que le *Big Data* est sa principale condition technique de possibilité. Dès lors, la production, l'organisation et le partage des données constituent son essence. Également, la production des données du projet est caractérisée par la nécessité de convertir en langage binaire des échantillons biologiques. Cela requiert des équipements et des équipes spécialisés dans ce processus en plus de fournir des données aux caractéristiques homogènes. Cette dernière contrainte fait en sorte qu'un mode de gestion impliquant des procédures uniformes pour l'ensemble des équipes doit être mis de l'avant pour que les données produites puissent s'intégrer dans la base de données génétique. Corolairement, les rôles et les

tâches sont divisés de façon à ce que chaque équipe effectue un travail différent sans pouvoir empiéter sur les tâches des autres membres du consortium.

Au fur et à mesure que les projets scientifiques prennent de l'expansion et que le financement afflue, les groupes de recherches se transforment en entreprise où les professeurs-chercheurs deviennent des administrateurs, adoptent des modes organisationnels entrepreneuriaux et recherchent activement du financement (Etzkowitz, 2003). Dans cet ordre d'idée, on constate que l'organisation et la division scientifique du travail au sein d'ENCODE s'effectue sous le mode du consortium où les tâches sont sectorisées.

Par exemple, le travail d'expansion des données génétiques est effectué par les équipes de *mapping* (FOA RFA-HG-16-002), qui constituent en quelque sorte les ouvriers de production du projet. Les groupes de *characterizing* (FOA RFA-HG-16-003) produisent des analyses et effectuent un travail de recherche et développement. Les équipes d'analyse et de développement informatique (FOA RFA-HG-16-004) responsable d'améliorer et de développer les méthodes d'analyse. Les équipes de coordination (FOA RFA-HG-16-005, FOA RFA-HG-16-006) agissent en tant que gestionnaire de projet, s'assurent de coordonner les groupes, d'effectuer un suivi et de classer et diffuser les données de l'encyclopédie génétique.

Cette division du travail – similaire à une entreprise de grande taille – est caractéristique de ce que l'on nomme les *big sciences*, ces supers projets dont le financement est élevé et l'organisation s'effectue en groupes ayant des tâches spécifiques, puisque celles-ci nécessitent une spécialisation accrue. D'ailleurs, le financement de la phase 4 d'ENCODE oscille entre 30 à 38 millions par années sur une période de 4 ans, pour un total avoisinant les 120 à 160 millions \$ (National Institutes of Health, 2017).

Évidemment, il n'est pas nécessaire que le *Big Data* soit mobilisé pour qu'un projet ait une envergure équivalente ou supérieure à ENCODE. Cependant, les contraintes

techniques d'ENCODE – conversion de données génétiques en langage binaire, taille du code génétique, données en libre accès – et son objectif de révéler sous langage binaire l'ensemble du *code de la vie* (génétique) implique un mode organisationnel s'apparentant à celui d'une entreprise. En ce qui concerne son mode organisationnel, le *Big Data* n'apparaît pas comme déterminant. Ce sont cependant les buts et l'envergure du projet, rendu techniquement possible par le *Big Data*, qui constituent le principal déterminant du mode d'organisation.

5.2 De l'*épistémè* à la *doxa* : cybernétique et société

Cette section s'intéresse à l'influence de l'*épistémè* cybernétique sur la *doxa*. Percer le noyau de la *doxa* post-moderne est une tâche dont on s'acquitte en revisitant le concept de réification sous l'angle de l'économie, puis sous celui de la science cybernétique. Aborder une version actualisée de ce concept sous l'angle du paradigme cybernétique permet de comprendre le passage de l'*épistémè* au statut de *doxa*.

Platon distingue l'*épistémè* de la *doxa*¹⁴, le premier concept fait référence aux formes de savoirs et de connaissances, tandis que second est considéré comme les croyances communes¹⁵ (Lafrance, 1982). Bien que des difficultés d'ordre proprement linguistique relatives au grec ancien et des incohérences dans l'œuvre du philosophe, il ressort une définition de la relation *épistémè-doxa* qui s'impose chez Platon. L'élève de Socrate émet à plusieurs égards une distinction d'ordre dichotomique dans cette relation.

¹⁴ δόξα en grec antique.

¹⁵ Pour une discussion étendue sur le couple *épistémè-doxa* dans l'œuvre de Platon, j'encourage le lecteur à consulter l'article : *Les fonctions de la doxa-épistémè dans les dialogues de Platon* (1982) de Lafrance. L'auteur aborde la thématique *épistémè-doxa*, non pas à partir des concepts contemporains, mais à partir d'éléments pré-socratique et socratique. Cette mesure permet de rendre justice à la pensée platonicienne en la contextualisant dans les débats de l'époque, inconnue aux distinctions contemporaines des débats sur la connaissance.

La philosophie pré-socratique traduit un pessimisme face à l'Anthropos – associant la *doxa* aux imperfections de la connaissance humaine – l'*épistémè* est associée à l'entéléchie de la connaissance divine. Le véridique est alors de nature divine et stable, tandis que la *doxa* n'est qu'apparence et volage. Platon confirme ce présupposé ancré dans la divinité et soutient que la *doxa* relève de l'ontologiquement faillible, tandis que l'*épistémè* relève de l'inaffabilité ; les deux étant irréconciliables. Il ajoute que le monde est instable dans le cadre de la *doxa* et immuable en ce qui a trait à l'*épistémè*. De Platon nous retenons une vive opposition entre les concepts de *doxa* et *épistémè*, les deux semblant irréconciliables.

En faisant un saut dans l'histoire jusqu'à la période des temps modernes, on constate que le terme *doxa* porte un bagage moins péjoratif dans les théories sociologiques contemporaines. Par exemple, Bourdieu décrit la *doxa* comme ce qui semble aller de soi, un accord tacitement accepté sur l'état du monde social et naturel pour une civilisation à une époque donnée (Bourdieu, 2015 [1972]). Loin d'être le résultat d'une histoire téléologique, la *doxa* est changeante et est un outil pour faire valoir une vision commune du monde.

Autrefois dominée par la religion – en tant qu'idéologie et moyen d'ériger du sens – la *doxa* est dorénavant propulsée par une nouvelle force, l'univers scientifique¹⁶. Les sciences offrent une nouvelle manière de comprendre le monde en plus d'être imperméable à la critique, se déployant sous le spectre de la neutralité – synonyme de *progrès* (McCormick, 1997). Les sciences et la technique – sous le paradigme cybernétique – sont le moteur de production de sens de la société post-moderne. Constellée de l'aura d'exclusivité à l'accession de la vérité, la technique se pose comme

¹⁶ On note que Marcuse place l'économie au centre d'un processus d'aplanissement du monde social pour le ragréer de façon unidimensionnelle où les conflits de classes semblent être caducs (Marcuse, 1968). Dans l'œuvre de Lukács, le concept de réification – au sens économique – se rapproche selon nous du d'une forme de *doxa* en intelligibilisant l'entièreté du monde social et naturel dans une pensée instrumentale (Lukács, 1984 [1960]). Sans minorer le rôle du mode de pensée économique dans la formation de la *doxa*, nous avons plutôt choisi d'étudier la piste portant sur le fait scientifique.

moteur du développement scientifique et vice-versa. Le processus de réification scientifique totale du monde est alors au mieux vu comme positif pour le développement de l'Humanité, au pire semble aller de soi sans qu'une quelconque attention critique y soit accordée.

Comme le mentionne Habermas dans son ouvrage *La technique et la science comme « idéologie »* publiée originellement en 1968, les sciences constituent une idéologie au sens sociologique du terme (Habermas, 2000 [1968]). En tant qu'idéologie, elles permettent de donner une représentation du monde, et ce, de façon plus ou moins inconsciente. Or, les sciences sont à l'accoutumée présentées dans une position d'opposition à l'idéologie, s'autojustifiant en tant que critique de celle-ci.

Notre réflexion se base principalement sur les deux postulats suivants : soit la primauté de l'idéologie scientifique comme moyen de formation de sens et de représentation du monde, puis le fondement cybernétique des techno-sciences mettant sur un dénominateur commun l'ensemble du monde socio-naturel en le convertissant en langage binaire. Les sciences post-modernes sont alors comprises comme processus idéologique de rationalisation du monde en tant que structure réifiante au profit de la mise en commun de la Nature & Société, puis du vivant & non-vivant. Alors que la technique du *Big Data* influence la science, comment cette vision se transpose-t-elle en tant que rapport idéologique?

5.2.1 De la réification cybernétique

Pour comprendre le passage de l'*épistémè* scientifique à la *doxa*, nous empruntons le chemin de la théorie de la réification. La réification scientifique (cybernétique) nous permet d'ouvrir une voie de passage pour lier *épistémè* et *doxa*. Il serait en effet imprudent de considérer que l'*épistémè* scientifique s'inscrit directement dans la *doxa*, sans en expliquer le procès. Pour expliquer ce processus, nous présentons initialement la théorie de la réification de Lukács, puisqu'elle est particulièrement développée quant

à ce sujet. Ensuite, nous expliquons les tenants et aboutissants de ce qu'on entend par le procès de réification post-moderne ou encore réification cybernétique.

Le concept de réification émane du terme latin *rei*, génitif de *res*, signifiant chose. À première vue, le concept est simplement entendu comme un processus de choséification. Georg Lukács – philosophe marxiste d'origine hongroise – théorise la réification comme l'objectivation de l'entièreté des rapports sociaux sous le joug de l'économie capitaliste. Dans le procès de réification capitaliste, la pensée marchande domine et transforme l'ensemble du monde naturel et social en les objectivant en fonction de la production, du profit, etc. (Lukács, 1984 [1960]).

La conscience de l'Anthropos étant désormais née sous le règne du capitalisme, celui-ci apparaît comme naturel et allant de soi, en tant que *doxa*. Le rapport à la société est alors transformé, le sujet pensant conçoit les rapports sociaux en fonction de production économique. Autrement dit, c'est la réduction de la substance de l'humanité, une choséification de la vie où l'être est dépersonnalisé au point où il n'est que rouage du capitalisme. Par exemple, le terme *ressource humaine* s'est imposé pour rendre compte de la constitution de l'individu au sein de l'entreprise, voire de la société. Considéré comme ressource dépersonnalisée, le sujet doit s'auto-interpréter en fonction de sa place dans l'appareil économique.

C'est d'ailleurs l'inversement noté par Arendt dans *La crise de la culture* (1964), où le statut du citoyen était – dans la Grèce antique – accordé à celui qui n'avait pas à travailler pour s'impliquer pleinement dans la vie démocratique. Dans le monde réifié économiquement, le statut de citoyen est mué en payeur de taxes et sa place dans la hiérarchie sociale – voire dans le *kosmos* – s'établit en fonction du travail, des revenus, etc.

On note que le concept de réification est habituellement mobilisé sous la forme précédemment exposée. Nous proposons ici un dérivé du concept de réification pour

comprendre la *doxa*. Le contexte économique et politique de l'œuvre de Lukács – caractérisé par un nouveau stade du capitalisme et d'organisation de la production fétichisant la marchandise – nécessitait une analyse en conséquence. Alors que la science et la technique sont désormais le principal moteur de production et d'idéologie, il est nécessaire de revisiter le concept de réification pour l'actualiser dans la société post-moderne ¹⁷.

Corolairement, il est impératif – pour distinguer les concepts de réification – de préciser qu'on traite de la réification scientifique post-moderne (cybernétique). Le procès de réification scientifique post-moderne est l'objectivation des catégories totales de la conscience en tant qu'objet relié au domaine de l'épistémologie cybernétique. Cette épistémologie se répercute alors dans l'ensemble de la conception du monde social et naturel. Le langage binaire en tant que norme communicationnelle et la conception machinique de la réalité deviennent un mode de compréhension et de représentation généralisé.

Du côté du corps humain ou animal, la réification cybernétique permet de réduire la morphologie – par exemple les organes – en machines ayant des fonctions spécifiques. L'article *Thoughts on the Human Body* (2010 [~1960]) est une publication *post-mortem* du chirurgien et chercheur médical John Ochsner, rédigée dans les années 1960. Cet article démontre les débuts de la représentation réifiée du corps humain. La naissance de l'ordinateur et du courant cybernétique ont certainement influencé l'auteur comme en témoigne les représentations du corps sous une forme mécanique : « When we arrive on this earth we are endowed with the most perfect, the most efficient, and the best constructed machine ever devised – our body » (Ochsner, 2010, p. 44), « The body's computer, the brain, is by far the most sophisticated, the finest constructed, the most

¹⁷ Pour des discussions plus poussées concernant la place de la technique dans le développement économique capitaliste, je recommande la lecture des textes suivants : Luke, T. (1980). « Technics » And Marx's Materialist Conception Of History. *The Indian Journal of Political Science*, 41(4), 695-728 ; Axelos, K. (1977). *Alienation, Praxis and Techne in the Thought of Karl Marx*.

efficient computer that has ever been or ever will be designed » (Ochsner, 2010, p. 44). Cette représentation du corps humain où le cerveau est le poste de commandement et les organes des machines ayant des tâches spécifiques permet la possibilité d'intelligibiliser l'Anthropos sous la forme d'un dispositif mécanique.

Un autre exemple de la représentation cybernétique du corps – cette fois plus proche de la communication binaire – est celui de la conception de la mémoire et de l'apprentissage. Les sciences de l'éducation et de l'apprentissage doivent dorénavant faire face aux mythes où l'individu est considéré comme un ordinateur et le savoir comme des données (Baillargeon, 2013). Le cerveau humain est alors compris comme un ordinateur et l'intelligence serait semblable et proportionnelle à la puissance de calcul d'un processeur. On note que les racines de cette conception remontent à l'invention des premiers ordinateurs :

In the 1940s got us especially confused. For more than half a century now, psychologists, linguists, neuroscientists and other experts on human behaviour have been asserting that the human brain works like a computer (Epstein, 2016).

La troisième représentation que nous souhaitons aborder émane de notre réflexion portant sur ENCODE. Évidemment, les sciences et la technique ne sont pas statiques, les développements techniques font progresser l'*épistémè* qui se répercute dans la *doxa*. Le domaine bio-informatique s'est imposé comme régime de vérité pour donner un sens à la vie. Conjuguée au paradigme cybernéticien, la bio-informatique offre une explication de la vie sous forme de code informatique à partir du code génétique. Or, les aspirations d'ENCODE de convertir l'entièreté de ce code sous une forme binaire ouvre la porte à une compréhension entièrement informatique de la vie. Cette perspective nous indéniablement la compréhension de la vie proportionnellement à la puissance informatique. Le *Big Data* est alors une étape technique catalysant l'idée d'un être purement informatique.

Ces trois représentations reposent sur le socle commun d'un procès de réification dont la transition s'effectue de l'*épistémè* cybernétique et se mue en tant que *doxa*. Cela s'explique par le fait que la post-modernité scientifique est basée sur le paradigme cybernéticien et que celui-ci avait déjà à son origine non pas un programme strictement scientifique, mais philosophique expansible à l'ensemble de la société (Wiener, 1950).

CONCLUSION

Ce mémoire a permis de documenter et saisir le rôle et une portion des transformations opérées par le *Big Data* dans le domaine scientifique, plus précisément dans le champ des sciences bio-informatiques. Notre objectif était de rendre compte des changements qui s'opèrent en science et dans la société sous le spectre du *Big Data*. En prenant soin d'expliquer que les changements techniques ne sont pas inéluctables et téléologiques, on comprend que leurs dénouements sont contingents à de nombreux facteurs et que les effets peuvent à la fois être directs ou indirects.

Notre problématique a permis d'expliquer les tenants et aboutissants du *Big Data* sous son aspect technique. Ainsi, nous avons présenté le *Big Data* sous une logique d'ensemble de 3V, rassemblant les catégories du volume, de la vitesse et de la variété. Le volume faisant référence à la capacité de production et de stockage des données numériques. Quant à elle, la vitesse est relative à la capacité de transmettre et de produire – quasi instantanément – une quantité d'information donnée. La variété comprend l'ensemble des différentes données composant un phénomène. La thématique de la valeur des données a été abordée pour expliquer qu'elles n'ont pas de valeur économique ou scientifique en soi. En effet, elles doivent s'inscrire dans un système complexe de *création* de la valeur en fonction de leur utilisation immédiate ou potentielle.

De plus, nous avons brièvement traité sur sujet de la gouvernamentalité algorithmique, s'inscrivant dans le contexte d'une perte de confiance envers les institutions politiques, journalistiques, etc. La confiance passe des institutions aux processus – les chiffres et algorithmes sont alors perçus comme neutre – la quantité de données étant proportionnelle à la bonne gestion du monde social.

Pour initialement joindre le sujet du *Big Data* et des sciences, nous avons abordé la thèse néo-empiriste portant sur le *Big Data* en science, tel que développé par Anderson dans son célèbre article : *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* (2008). *Grosso modo* la thèse néo-empiriste est basée sur 4 postulats. *Primo*, qu'il est possible de cerner l'ensemble des données relatives à un phénomène. *Deuxio*, que la mobilisation théorique est néfaste pour interpréter les données, car elles parlent d'elles-mêmes. *Tertio*, qu'il faut écarter la manipulation humaine au maximum pour arriver à atteindre l'objectivité et l'hyperréel. *Quarto*, que l'utilisation de concepts et de théories devient désuet, il y aurait effacement des disciplines, seul le traitement informatique des données serait utile pour déceler le *réel*.

Le problème est que ces postulats, présentés par Anderson en 2008, ne concordent pas avec la réalité du processus scientifique. Cela nous a mené à explorer l'approche *data-driven science*, ce mode d'investigation scientifique où les données sont l'étincelle menant à des recherches cadrées théoriquement. Ces éléments exposés, nous avons opté d'explorer l'influence des *Big Data* sur le domaine scientifique en étudiant le champ des sciences bio-informatiques.

Notre cadre théorique a croisé plusieurs concepts provenant de l'étude des sciences et de la technique sous ses branches à la fois philosophique et sociologique. Par exemple, les concepts de déterminisme technique et de neutralité technique ont été présentés pour éviter de tomber dans le piège où la technique serait considérée en tant que téléologie ou sans aucune essence. À cet égard, nous avons favorisé un point de vue où la technique est une condition de possibilité.

Pour contextualiser plus largement les changements opérés dans le domaine scientifique et sociétal, nous avons traité du sujet de la modernité scientifique et de la post-modernité. Le développement technique et scientifique ayant remplacé la religion et partiellement l'économie comme moyen de donner un sens au monde, l'étudier permet de jeter un regard sur une conception du monde dont la base est scientifique.

Le paradigme cybernéticien s'est rapidement imposé dans l'ensemble du processus socio-technique et en est venu à être une caractéristique intrinsèque de la post-modernité. Cette transformation est d'autant plus insidieuse qu'elle semble aller de soi, sous le couvert de la thèse techno-déterministe où le progrès évoluerait de façon *naturelle* vers un objectif prédéterminé. Intimement liée au concept de progrès unilinéaire, l'histoire post-moderne est désormais perçue comme une téléologie dénuée de contingence (Arendt, 2000 [1964]). Le pivot historique est alors l'éternel présent dont l'histoire émanerait sans cesse. La double infinité du passé et du futur n'est pensée qu'à partir du présent. C'est ainsi que le passé est pensé à partir du présent – et non plus l'inverse – masquant la contingence du développement socio-technique. En ce sens, bien que la généralisation de l'informatique et du paradigme cybernéticien semble aller de soi, c'est tout un processus impliquant des changements dans les racines de l'*épistémè* et de la *doxa*.

Pour étudier le phénomène du *Big Data* en science, nous avons arrêté notre choix sur le domaine des sciences bio-informatiques, celui-ci succédant à la biologie moléculaire. Pour observer les changements s'opérant sous le spectre du *Big Data*, notre cas de recherche sélectionné est le projet ENCODE. En tant que figure de proue de l'utilisation des *Big Data* en science, autant par la durée du projet que sa structure, il constitue l'objet de recherche le plus complet pour notre étude. C'est grâce à l'analyse de son article de référence, de sa politique de diffusion des données et de documents de subvention nous avons produit cette recherche.

Réponse aux questions de recherches

Du début à la fin, cette recherche a été structurée de façon à étudier à la fois le rôle du *Big Data* en science, puis les sciences dans la société. Dès lors, notre recherche a naturellement été divisée en deux sections complémentaires. La première section traite

de l'influence des *Big Data* sur l'épistémè scientifique, tandis que la seconde porte sur les changements opérés par les sciences *Big Data* sur la *doxa*.

Dans le cadre de notre première sous question de recherche, notre interrogation s'est posée sur l'influence du *Big Data* en science. Pour y répondre, notre analyse s'est divisée en 5 sections complémentaires. En tant que phénomène technique récent, le *Big Data* est considéré comme supérieur – autant au niveau de la vitesse que du volume de données – face aux *anciennes* méthodes scientifiques. Le programme scientifique cybernétique – en tant que structure ontologique – n'échappe pas à la relation dialectique science/technique. De pair avec la modulation à la hausse des limitations de nature technique, le *Big Data* amplifie et réalise les aspirations du paradigme cybernéticien en science sous de multiples angles.

Premièrement, nous avons étudié la forme de validité des savoirs dans les sciences à partir du projet ENCODE. Cette section a initialement été inspirée de notre lecture de Lyotard où il mentionne que dans la société post-moderne :

[...] la nature du savoir ne reste pas intacte. Il ne peut passer dans les nouveaux canaux, et devenir opérationnel, que si la connaissance peut être traduite en quantité d'information. On peut donc en tirer la prévision que tout ce qui dans le savoir constitué n'est pas ainsi traduisible sera délaissé, et que l'orientation des recherches nouvelles se subordonnera à la condition de traduisibilité des résultats éventuels en langage de machine (Lyotard, 1979, p. 13).

On peut avancer que la prédiction de Lyotard s'est partiellement réalisée au sens où la validité du savoir s'est en effet déplacée d'un mode de compréhension qualitatif à quantitatif. En se présentant en tant qu'encyclopédie, le projet ENCODE dévoile ses prétentions au remplacement de ceux-ci par des bases de données. En ce sens, le savoir deviendrait quantitatif et proportionnel aux données disponibles ; au détriment de la compréhension par concept. On remarque par contre que le projet ENCODE est avant tout un producteur de donnée et que son mode d'analyse demeure restreint.

Deuxièmement, nous avons convenu de nous pencher sur le socle épistémologique d'ENCODE. Les sciences de la vie sont désormais techniquement tributaires des développements de l'informatique et de l'électronique. Cette limitation technique est expliquée parce que la vie est désormais considérée sous forme de code informatique – par exemple le code génétique. Autrement dit, le sujet et la vie sont choséifiés sous le paradigme communicationnel qui considère que l'Être et la vie sont encodés sous le langage binaire. Le *Big Data* rend techniquement possible la réalisation de base de données ayant la possibilité d'encoder l'entièreté du code génétique.

Troisièmement, nous avons étudié le mode d'analyse et de compréhension du projet. Comme l'objectif du consortium est principalement de produire des données, il mobilise un mode de compréhension basé sur des corrélations entre les variables. Comme les laboratoires produisant les données n'ont pas des ressources illimitées, la méthode corrélatrice permet de circonscrire les données à produire en mettant de l'avant le potentiel d'interaction entre les variables génétiques.

Quatrièmement, notre attention s'est posée sur l'aspiration du projet quant à l'exhaustivité des données. Le volume de données – particulièrement du côté de la production – semble l'élément le plus important en comparaison à la variété et la vitesse. En ayant l'aspiration de produire une base de données complète portant sur les données génétiques, le projet ENCODE met de l'avant l'idée qu'il est possible d'arriver à atteindre l'exhaustivité des données pertinentes sur un sujet donné.

Le dernier élément porte sur le processus organisationnel du projet ENCODE. D'un programme scientifique à l'autre, le *Big Data* peut occuper une place centrale ou non. Concernant notre cas d'étude, le *Big Data* est primordial dans la mesure où il en est une condition de possibilité technique. Il rend possibles les aspirations de production, de stockage et de partage d'une massive base de données. Or, tout comme le travail à la chaîne implique intrinsèquement un processus d'organisation sociale en fonction du temps quantitatif mesuré, les contraintes socio-techniques d'ENCODE tendent à

expliquer son organisation. Les contraintes relatives à la taille et la complexité du projet ENCODE sont une cause directe de son mode organisationnel où le travail est méthodiquement divisé de façon similaire à une entreprise.

Dans le cadre de notre seconde sous-question de recherche, notre attention s'est posée sur l'influence des sciences *Big Data* sur la *doxa*. Pour ce faire, nous avons choisi d'ouvrir une voie de passage sur la base de deux postulats. Le premier est que les sciences et la technique constituent une idéologie, un moyen de donner un sens au monde social et naturel. En ce sens, l'*épistémè* scientifique post-moderne se transpose dans la *doxa*. Deuxièmement, que l'idéologie scientifique est marquée par le courant cybernéticien. Ainsi, l'idéologie scientifique cybernétique se répercute – plus ou moins consciemment – sur la société.

Or, si ces deux postulats vont de pair, il est nécessaire de les préciser en expliquant leurs tenants et aboutissants. Les sciences *Big Data* sont de nature technique, mais également phénoménologique au sens où l'objet est avant tout filtré par une série de concepts et structure mentale. Dès lors, se questionner sur l'influence des sciences *Big Data* sur la société impose de porter notre regard non pas seulement sur un objet technique, mais sur les structures mentales sous-jacentes au phénomène. Comme nous l'avons mentionné, l'élément systémique abstrait catalysant les sciences *Big Data* est le paradigme cybernétique.

Pour décrire cette influence sur la société, nous avons considéré le paradigme cybernétique sous l'angle de l'*épistémè* et la société en tant que *doxa*. En mettant l'accent sur le passage de l'*épistémè* à la *doxa*, on propose une version actualisée du concept de réification basé sur le paradigme cybernétique. Dès lors, l'influence du processus de réification cybernétique sur la *doxa* est la considération de l'ensemble des processus naturels et sociaux sous le seul angle du méta-langage binaire.

Ce procès d'arasement du monde social et naturel est caractéristique de la post-modernité où ce dualisme hypostasié est remplacé par le méta-langage binaire comme forme de compréhension hégémonique. Comme le méta-langage binaire permet de poser le monde sur un dénominateur commun, celui-ci est choséifié sous le langage binaire et une compréhension mécanique, puis informatique de la vie. La vie ne serait plus qu'un code informatique dont la compréhension et le traitement seraient proportionnels à la puissance informatique.

Limites de la recherche

Nous avons évoqué en détail les circonstances portant sur l'étude des *Big Data* en science, nos questionnements de recherche et nos résultats. Il est maintenant temps de se pencher sur les limites de notre processus de recherche. D'abord, on soulève que notre étude s'est intéressée à un seul projet impliquant le *Big Data*, et ce, dans un domaine scientifique spécifique : les sciences bio-informatiques. Les restrictions imposées par l'exercice de rédaction d'un mémoire de recherche nous ont poussés à sélectionner un projet qui nous semble représentatif des développements des *Big Data* en science. Or, le *Big Data* se déploie certainement à des échelles variées par rapport aux domaines scientifiques et des variations sont possibles à l'intérieur même des champs de recherches. Ainsi, étudier les tenants et aboutissants du *Big Data* dans d'autres domaines permettrait de jeter un regard plus large sur ce phénomène socio-technique.

Au niveau théorique, nous avons favorisé un cadre conceptuel nous laissant la liberté d'interpréter nos résultats à la fois à la vue du développement socio-technique du *Big Data* sur les sciences, mais aussi sur la société. Or, étudier un seul cas pour contribuer à la compréhension de l'ensemble du processus post-moderne en science demeure

partiel, bien que pouvant contribuer à mieux comprendre le pan des sciences bio-informatiques.

Une autre limite de notre recherche que les changements socio-techniques peuvent parfois s'établir sur une très courte période ou bien se cristalliser et demeurer hégémoniques pour une période étendue. En ce qui concerne le cas de la technique dans le domaine des communications, son évolution rapide ouvre de nouveaux champs d'études au fur et à mesure de leur développement. Dès lors, le domaine des communications semble substantiellement sensible face aux changements socio-techniques s'opérant, plus souvent qu'autrement, à la hâte.

Piste de recherche : La gouvernance par les sciences de la communication? Du gouvernement à la gouvernementalité hyperréelle

Ce serait faire preuve de cécité que de ne pas constater une relation quasi incestueuse entre le domaine d'étude des communications et les nouvelles technologies. Comme nous l'avons mentionné précédemment, la technique est avant tout une condition de possibilité et ne guide pas unilatéralement le développement social, elle peut cependant poser les jalons de nouvelles façons de faire, voire de penser. L'adoption d'une nouvelle technologie implique nécessairement des effets extrinsèque et intrinsèque à celle-ci. Cependant, les effets à long terme sont habituellement balayés du revers de la main au profit d'une vision s'alignant avec des intérêts économiques ou encore en se présentant comme la succession logique du développement socio-technique.

On rappelle que les sciences et la technique s'inscrivent dans une relation dialectique qui se répercute dans la *doxa*. Étudier un élément technique dans une perspective sociologique revient alors à s'intéresser à la société dans son ensemble. Or, la période idéale pour procéder à l'étude d'un élément socio-technique est lorsqu'il y a *controverse*, c'est-à-dire qu'une transformation technique s'opère ou est sur le point de

le faire. Il est ainsi possible de comprendre pourquoi un choix a été privilégié au détriment d'un autre et d'assister aux transformations quand la technique et la science sont *en action* ¹⁸.

Plus largement, le *Big Data* catalyse l'aspiration à la réification cybernétique – sous code informatique binaire – de l'ensemble du monde social et naturel sous un dénominateur commun. Il permet d'amplifier autant le volume, la vitesse et la variété des données produites, stockées, puis analysées. Or, les sciences *Big Data* donc sont un moyen d'augmenter les potentialités de contrôle et d'emprise sur le monde en faveur d'une hyperréalité illimitée. Les sciences *Big Data* sont donc un point d'expansion significatif quant aux possibilités d'un état de gouvernementalité se prolongeant jusqu'à la composition des cellules. Les sciences et la technique se muent donc en formidable moyen d'amplification du pouvoir de contrôle.

L'optimisme envers la gouvernementalité algorithmique est à la fois basé sur sa présumée capacité à prendre la décision optimale, et ce, en mobilisant un imaginaire où la machine bienveillante se préoccupe en permanence d'accomplir sa responsabilité. En réifiant le vivant à l'échelle de sa composition bio-chimique sous le paradigme cybernétique, la gouvernance par la machine est catalysée par l'idéal d'omniscience en prenant en compte des données jusqu'au niveau cellulaire. Bien qu'encore loin d'être concrétisés dans le domaine cellulaire, des projets scientifiques de gouvernementalité algorithmiques sont mis en place en ce qui a trait à des analyses de masse en temps réel concernant des modèles neurocomportementaux.

Par exemple, le gouvernement américain jongle avec l'idée de créer une nouvelle branche de recherche nommée *Health Advanced Research Projects Agency* (HARPA), celle-ci serait l'homologue de la *Defense Advanced Research Projects Agency*

¹⁸ Pour de plus amples informations sur la pertinence de l'étude des sciences et de la technique *en action*, nous recommandons le classique *La science en action* (1989) de Bruno Latour.

(DARPA). Cette agence aurait comme projet de développer – en utilisant des données provenant des appareils intelligents – un programme d’intelligence artificielle pour détecter des signes s’inscrivant dans des modèles neurocomportementaux pour prévoir quels individus sont à risque d’effectuer des actes violents, comme des fusillades, et ce, en temps réel (Alemany, 2019 ; Ehrenkranz, 2019).

Or, ce mode de compréhension basé sur les données ne tient pas en compte de théorie sociologique ou politique concernant le problème de la violence par armes à feu aux États-Unis. Cette méthode consiste en une production maximale de donnée incluant un plan de surveillance de masse dont la raison n’est plus nécessairement la sécurité, mais la bonne gouvernance et le contrôle.

BIBLIOGRAPHIE

Alberts, B. (2012). The End of « Small Science »? *Science*, 337(6102), 1583-1583. doi: 10.1126/science.1230529

Alemany, J. (2019, 22 août). White House considers new project seeking links between mental health and violent behavior. *Washington Post*, section Politics. Récupéré de <https://www.washingtonpost.com/politics/2019/08/22/white-house-considers-new-project-seeking-links-between-mental-health-violent-behavior/>

Alvarez León, L. F. (2016). Property regimes and the commodification of geographic information: An examination of Google Street View. *Big Data & Society*, 3(2), 2053951716637885. doi: 10.1177/2053951716637885

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Dans *WIRED*. Récupéré de <https://www.wired.com/2008/06/pb-theory/>

Arendt, H. (2000). *La crise de la culture: huit exercices de pensée politique*. Paris : Gallimard.

Armbrust, M., Stoica, I., Zaharia, M., Fox, A., Griffith, R., Joseph, A. D., ... Rabkin, A. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50. doi: 10.1145/1721654.1721672

Baillargeon, N. (2013). *Légendes pédagogiques: l'autodéfense intellectuelle en éducation*. Montréal (Québec) : Éditions Poètes de brousse.

Bain, C., Ransom, E. et Worosz, M. R. (2011). Constructing Credibility: Using Technoscience to Legitimate Strategies in Agrifood Governance. *Journal of Rural Social Sciences*, 25(3), 160-192.

- Baudrillard, J. (1981). *Simulacres et simulation*. Paris : Galilée.
- Bhaskar, R., Danermark, B. et Price, L. (2017). *Interdisciplinarity and wellbeing: a critical realist general theory of interdisciplinarity*. Abingdon, Oxon ; New York, NY : Routledge.
- Birney, E. (2012). The making of ENCODE: lessons for big-data projects. *Nature*, 489(7414), 49–51.
- Bourdieu, P. (2015). *Esquisse d'une théorie de la pratique: précédé de Trois études d'ethnologie kabyle*. Paris : Éditions du Seuil.
- Boyle, R. (2013). The Drama Over Project Encode, And Why Big Science And Small Science Are Different. *Popular science*. Récupéré de <https://www.popsci.com/science/article/2013-02/fracas-over-project-encode-and-why-big-science-and-small-science-are-different>
- Caidi, N., Shankar, K., Dalbello, M. et Froehlich, T. J. (2005). How neutral can technology be? *Proceedings of the Association for Information Science and Technology*, 42(1).
- Canali, S. (2016). Big Data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS. *Big Data & Society*, 3(2), 2053951716669530. doi: 10.1177/2053951716669530
- Charter, C., Fellows, W., Fellows, D., women Fellows, D. et Fellows, D. H. (2007). *Jim Gray on eScience: A Transformed Scientific Method*. Récupéré de <http://www.ias.ac.in/describe/article/reso/021/08/0749-0763>
- Chassay, J. F. (2006). La confrontation des genres. Dans *Culture et technoscience : des enjeux du sens à la culture. Approche d'une logique multidisciplinaire*. Québec : Les Presses de l'Université Laval.
- Clain, O. (1989). Sur la science contemporaine. *Société, Hiver*.
- Clarke Adele E. (2014). Biomedicalization. *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society*. doi: 10.1002/9781118410868.wbehibs083

- Cohen, I. B. (2005). *The triumph of numbers: how counting shaped modern life*. Récupéré de <https://www.overdrive.com/search?q=3FFAF16A-CEEA-4E03-8C3A-C204323F3719>
- Crawford, K. (2013). The Hidden Biases in Big Data. *Harvard Business Review*. Récupéré de <https://hbr.org/2013/04/the-hidden-biases-in-big-data>
- Daston, L. et Galison, P. (1992). The Image of Objectivity. *Representations*, (40), 81-128. doi: 10.2307/2928741
- Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., ... Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, 46(D1), D794-D801. doi: 10.1093/nar/gkx1081
- De Vries, G., Courtois-l'Heureux, F. et Pignarre, P. (2018). *Bruno Latour: une introduction*. (s. l. : n. é.).
- Dean, J. (2016). Big Data: Accumulation and Enclosure. *Theory & Event*, 19(3). Récupéré de <https://muse.jhu.edu/article/623988>
- Delgado, A. (2016). *Technoscience and citizenship: ethics and governance in the digital society*. New York, NY : Springer Berlin Heidelberg.
- Ecker, J. R., Bickmore, W. A., Barroso, I., Pritchard, J. K., Gilad, Y. et Segal, E. (2012). Genomics: ENCODE explained. *Nature*, 489(7414), 52–55.
- Ehl, C. (2018, 14 janvier). Data — the Fuel for Artificial Intelligence. *Christian Ehl*. Récupéré de <https://medium.com/@cehl/data-the-fuel-for-artificial-intelligence-ed90bf141372>
- Ehrenkranz, M. (2019). The Plan to Use Fitbit Data to Stop Mass Shootings Is One of the Scariest Proposals Yet. Dans *Gizmodo*. Récupéré de <https://gizmodo.com/the-plan-to-use-fitbit-data-to-stop-mass-shootings-is-o-1837710691>
- Eisenstein, E. L. (2005). *The Printing Revolution in Early Modern Europe*. (s. l.) : Cambridge University Press.

- ENCODE. (2014). *Data Policy*. Récupéré de <http://genome.ucsc.edu/encode/terms.html>
- ENCODE. (2018). *Access to ENCODE data – ENCODE*. Récupéré de <https://www.encodeproject.org/about/data-access/>
- ENCODE Project Consortium. (2012). An Integrated Encyclopedia Of DNA Elements In The Human Genome. *Nature*, 489(7414), 57-74. doi: 10.1038/nature11247
- ENCODE Project Consortium. (2014). *ENCODE Consortium Data Release Policy*. Récupéré de <https://www.genome.gov/27528022/encode-consortium-data-release-policy/>
- ENCODE Project Consortium. (2019a). *Citing ENCODE (Data Use Policy)*. Récupéré de <https://www.encodeproject.org/help/citing-encode/>
- ENCODE Project Consortium. (2019b). *Project Overview – ENCODE*. Récupéré de <https://www.encodeproject.org/about/contributors/>
- Epstein, R. (2016, 18 mai). Your brain does not process information and it is not a computer – Robert Epstein | Aeon Essays. *Aeon*. Récupéré de <https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer>
- Etzkowitz, H. (2003). Research groups as ‘quasi-firms’: the invention of the entrepreneurial university. *Research Policy*, 32, 109-121. doi: 10.1016/s0048-7333(02)00009-4
- Feenberg, A. (2013). L’anthropologie et la question de la Nature. Réflexions sur L’Écologie des autres, de Philippe Descola. *Revue du MAUSS*, (2), 105–118.
- Feenberg, A., Dassas, V., Callon, M. et Arnaq, I. (2014). *Pour une théorie critique de la technique*. Montréal (Québec) : Lux.
- Forman, P. (2007). The Primacy of Science in Modernity, of Technology in Postmodernity, and of Ideology in the History of Technology. *History and Technology*, 23(1-2), 1-152. doi: 10.1080/07341510601092191

- Foucault, M. (1966). *Les mots et les choses: une archéologie des sciences humaines*. Paris : Gallimard.
- Freitag, M. (2002). *L'oubli de la société: pour une théorie critique de la postmodernité*. Québec : Presses de l'Université Laval.
- Gagnon, Y.-C. (2012). *L'étude de cas comme méthode de recherche* (2nd ed). Québec : Presses de l'Université du Québec.
- Garvey, W. D. (1979). *Communication, the essence of science: facilitating information exchange among librarians, scientists, engineers, and students*. Oxford ; New York : Pergamon Press.
- George, É. et Granjon, F. (dir.). (2008). *Critiques de la société de l'information*. Paris : L'Harmattan.
- Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211-221. doi: 10.1007/s10708-007-9111-y
- Gordon, K. (2013). What is Big Data? *ITNOW*, 55(3), 12-13. doi: 10.1093/itnow/bwt037
- Grant, E. (2012). The promise of big data. *News*, 2012(Spring). Récupéré de <https://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/>
- Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A. et Elhaik, E. (2013). On the Immortality of Television Sets: « Function » in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biology and Evolution*, 5(3), 578-590. doi: 10.1093/gbe/evt028
- Gray, E., Jennings, W., Farrall, S. et Hay, C. (2015). Small Big Data: Using multiple data-sets to explore unfolding social and economic change. *Big Data & Society*, 2(1), 2053951715589418. doi: 10.1177/2053951715589418
- Habermas, J. (2000). *La technique et la science comme « idéologie »*. Paris : Gallimard.
- Harvey, D. (2006). *The limits to capital* (New and fully updated ed). London ; New York : Verso.

- Hey, A. J. G. (dir.). (2009). *The fourth paradigm: data-intensive scientific discovery*. Redmond, Washington : Microsoft Research.
- Huff, T. E. (2003). *The rise of early modern science: Islam, China, and the West* (2nd ed). Cambridge, U.K. ; New York : Cambridge University Press.
- Joly, C. (dir.). (2013). *La technoscience et ses enjeux: les transformations théoriques et empiriques de la science et du risque*. Montréal : Éditions libres du Carré rouge.
- Kay, L. E. (2000). *Who wrote the book of life? A history of the genetic code*. Stanford, Calif : Stanford Univ. Press.
- Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 2053951714528481. doi: 10.1177/2053951714528481
- Kitchin, R. et Lauriault, T. P. (2014). *Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work*. Récupéré de <http://scihub.cc/https://papers.ssrn.com/abstract=2474112>
- Kitchin, R. et McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1), 2053951716631130. doi: 10.1177/2053951716631130
- Kolata, G. (2012, 5 septembre). Far From 'Junk,' DNA Dark Matter Proves Crucial to Health. *The New York Times*, section Science. Récupéré de <https://www.nytimes.com/2012/09/06/science/far-from-junk-dna-dark-matter-proves-crucial-to-health.html>
- Kuhn, T. S. (1983). *La structure des révolutions scientifiques*. Paris : Flammarion.
- Lafleur, S. (2017). Penser la gouvernementalité après Foucault. *Politique et Sociétés*, 36(3), 141-159. doi: <https://doi.org/10.7202/1042239ar>
- Lafontaine, C. (2004). *L'empire cybernétique: des machines à penser à la pensée machine*. Paris : Seuil.
- Lafrance, Y. (1982). Les fonctions de la doxa-épistémè dans les dialogues de Platon. *Laval théologique et philosophique*, 38(2), 115. doi: 10.7202/705925ar

- Larousse, É. (2019). *Définitions : encyclopédie - Dictionnaire de français Larousse*. Récupéré le 23 mai 2019 de <https://www.larousse.fr/dictionnaires/francais/encyclopedie/29286>
- Latour, B. (1988). *The pasteurization of France*. Cambridge, Mass : Harvard University Press.
- Latour, B. (1999). *Pandora's hope: essays on the reality of science studies*. Cambridge, Mass : Harvard University Press.
- Latour, B. (2005). *Nous n'avons jamais été modernes: essai d'anthropologie symétrique*. Paris : Éditions La Découverte.
- Latour, B. et Biezunski, M. (2010). *La science en action: introduction à la sociologie des sciences*. Paris : Découverte / Poche.
- Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society*, 1(1), 2053951714534395. doi: 10.1177/2053951714534395
- Liddell, H. G. et Scott, R. (1940). *A Greek-English Lexicon : τέχνη*. Récupéré de <http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0057:entry=te/xnh>
- Logan, R. K. (2004). *The alphabet effect: a media ecology understanding of the making of Western civilization*. Cresskill, N.J : Hampton Press.
- Lowrie, I. (2017). Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data & Society*, 4(1), 2053951717700925. doi: 10.1177/2053951717700925
- Lukács, G. (1984). *Histoire et conscience de classe: essais de dialectique Marxiste*. Paris : Les Éditions de Minuit.
- Lyotard, J.-F. (1979). *La condition postmoderne: rapport sur le savoir*. Paris : Éditions de Minuit.
- Marcuse, H. (1964). *L'homme unidimensionnel: essai sur l'idéologie de la société avancée*. Paris : Ed de Minuit.

- Marcuse, H. (1968). *L'homme unidimensionnel. Essai sur l'idéologie de la société industrielle avancée*. Paris : Minuit.
- Marx, V. (2013, 12 juin). Biology: The big challenges of big data [News]. Dans *Nature*. doi: 10.1038/498255a
- Mathisen, B. M., Wienhofen, L. et Roman, D. (2015). Empirical Big Data Research: A Systematic Literature Mapping. *arXiv:1509.03045*. Récupéré de <http://arxiv.org/abs/1509.03045>
- Matteson, A. (2018). The Concept of Datafication; Definition & Examples. Récupéré de <https://www.datasciencecentral.com/profiles/blogs/the-concept-of-datafication-definition-amp-examples>
- Mayer-Schönberger, V. et Cukier, K. (2014). *Big data: a revolution that will transform how we live, work, and think* (First Mariner Books edition). Boston : Mariner Books, Houghton Mifflin Harcourt.
- McAfee, A., Brynjolfsson, E. et Davenport, T. (2012). Big data : the management revolution. *Harvard Business Review*, 90(10), 60-68.
- McCormick, J. P. (1997). *Carl Schmitt's Critique of Liberalism: Against Politics as Technology*. (s. l.) : Cambridge University Press.
- Merton, R. K. (1968). *Social theory and social structure*. New York, NY : Free Press.
- Michaud, T. (2017). De la science-fiction à l'innovation technoscientifique : le cas des casques de réalité virtuelle. *Innovations*, n° 52(1), 43-61.
- Miller, H. J. (2010). The Data Avalanche Is Here. Shouldn't We Be Digging? *Journal of Regional Science*, 50(1), 181-201. doi: 10.1111/j.1467-9787.2009.00641.x
- Mirowski, P. (2018). The future(s) of open science. *Social Studies of Science*, 48(2), 171-203. doi: 10.1177/0306312718772086

- National Human Genome Research Institute. (2017). 2017 Release: NIH to expand critical catalog for genomics research. Dans *National Human Genome Research Institute (NHGRI)*. Récupéré de <https://www.genome.gov/27567592/2017-Release-NIH-to-expand-critical-catalog-for-genomics-research>
- National Institutes of Health. (2016a). *FOA RFA-HG-16-002: Expanding the Encyclopedia of DNA Elements (ENCODE) in the Human and Mouse (UM1)*. Récupéré de <https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-16-002.html>
- National Institutes of Health. (2016b). *FOA RFA-HG-16-003: Characterizing the Functional Elements in the Encyclopedia of DNA Elements (ENCODE) Catalog (UM1)*. Récupéré de <https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-16-003.html>
- National Institutes of Health. (2016c). *FOA RFA-HG-16-004: Computational Analysis of the Encyclopedia of DNA Elements (ENCODE) Data (U01)*. Récupéré de <https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-16-004.html>
- National Institutes of Health. (2016d). *FOA RFA-HG-16-005: ENCODE Data Coordinating Center (U24)*. Récupéré de <https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-16-005.html>
- National Institutes of Health. (2016e). *FOA RFA-HG-16-006: ENCODE Data Analysis Center (U24)*. Récupéré de <https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-16-006.html>
- National Institutes of Health. (2017). *2017 News Release: NIH to expand critical catalog for genomics research - National Human Genome Research Institute (NHGRI)*. Récupéré de <https://www.genome.gov/27567592/2017-release-nih-to-expand-critical-catalog-for-genomics-research/>
- N'Da, P. (2015). *Recherche et méthodologie en sciences sociales et humaines: réussir sa thèse, son mémoire de master ou professionnel, et son article*. Récupéré de <http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1275916>
- Ochsner, A. (2010). Thoughts on the Human Body. *The Ochsner Journal*, 10(1), 44-47.

- Pagès, C. (2011). La fin des grands récits d'émancipation. *Philosophies*, 103-131.
- Patel, N. et Pilgrim, D. (2018). Psychologists and torture: critical realism as a resource for analysis and training. *Journal of Critical Realism*, 17(2), 176-191. doi: 10.1080/14767430.2018.1430975
- Porter, T. M. (1995). *Trust in numbers: the pursuit of objectivity in science and public life*. Princeton, N.J : Princeton University Press.
- Price, L. et Martin, L. (2018). Introduction to the special issue: applied critical realism in the social sciences. *Journal of Critical Realism*, 17(2), 89-96. doi: 10.1080/14767430.2018.1468148
- Rieder, G. et Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society*, 3(1), 2053951716649398. doi: 10.1177/2053951716649398
- Roderick, L. (2014). Discipline and power in the digital age: The case of the US consumer data broker industry. *Critical Sociology*, 40(5), 729-746.
- Rouvroy, A. et Berns, T. (2013). Gouvernamentalité algorithmique et perspectives d'émancipation, Faced with algorithmic governmentality. *Réseaux*, (177), 163-196. doi: 10.3917/res.177.0163
- Rouvroy, A. et Stiegler, B. (2015). Le régime de vérité numérique. De la gouvernamentalité algorithmique à un nouvel État de droit. *Socio. La nouvelle revue des sciences sociales*, (4), 113-140. doi: 10.4000/socio.1251
- Ruppert, E., Isin, E. et Bigo, D. (2017). Data politics. *Big Data & Society*, 4(2), 2053951717717749. doi: 10.1177/2053951717717749
- Sagiroglu, S. et Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42-47. doi: 10.1109/CTS.2013.6567202
- SCOPUS. (2019). Metric Details : An integrated encyclopedia of DNA elements in the human genome (2012). Dans *Scopus Metrics*. Récupéré de <https://www.scopus.com/record/pubmetrics.uri?eid=2-s2.0-84865790047>

- Séguin, È. (2012). La technoscience et le progrès du pouvoir [ACFAS]. *Chronique : La science c'est politique*. Récupéré de <http://www.acfas.ca/publications/decouvrir/2012/05/technoscience-progres-pouvoir>
- Sfez, L. (1990). *Critique de la communication* (Nouv. éd., entièrement réf. et augm). Paris : Seuil.
- Shapin, S. (2007). *A social history of truth: civility and science in seventeenth-century England* (4. print). Chicago : Univ. of Chicago Press.
- Singer, E. (2013). Biology's Big Problem: There's Too Much Data to Handle. Dans *WIRED*. Récupéré de <https://www.wired.com/2013/10/big-data-biology/>
- Smith, O. (2013, 1^{er} avril). Facebook terms and conditions: why you don't own your online life, section Technology. Récupéré de <http://www.telegraph.co.uk/technology/social-media/9780565/Facebook-terms-and-conditions-why-you-dont-own-your-online-life.html>
- Struijs, P., Braaksma, B. et Daas, P. J. (2014). Official statistics and Big Data. *Big Data & Society*, 1(1), 2053951714538417. doi: 10.1177/2053951714538417
- Thatcher, J., O'Sullivan, D. et Mahmoudi, D. (2015, 30 décembre). *Data Colonialism Through Accumulation by Dispossession: New Metaphors for Daily Data* [SSRN Scholarly Paper](ID 2709498). Rochester, NY : Social Science Research Network. Récupéré de [papers.ssrn.com](https://papers.ssrn.com/abstract=2709498) : <https://papers.ssrn.com/abstract=2709498>
- Venturini, T., Baya Laffite, N., Cointet, J.-P., Gray, I., Zabban, V. et De Pryck, K. (2014). Three maps and three misunderstandings: A digital mapping of climate diplomacy. *Big Data & Society*, 1(2). doi: 10.1177/2053951714543804
- Wagstaff, K. (2012). You'd Need 76 Work Days to Read All Your Privacy Policies Each Year. *Time*. Récupéré de <http://techland.time.com/2012/03/06/you-d-need-76-work-days-to-read-all-your-privacy-policies-each-year/>
- Wiener, N. (1950). *The Human Use of Human Beings: Cybernetics and Society*. University of Michigan : Houghton Mifflin.

- Wiener, N. (2014). *La cybernétique : information et régulation dans le vivant et la machine*. Paris : Seuil.
- Wiener, N. et Rosenblueth, A. (1946). The mathematical formulation of the problem of conduction of impulses in a network of connected excitable elements, specifically in cardiac muscle. *Archivos Del Instituto De Cardiologia De Mexico*, 16(3), 205-265.
- Winner, L. (2002). *La baleine et le réacteur: à la recherche de limites au temps de la haute technologie*. Paris : Descartes & Cie.
- Woot, P. (2005). *Should Prometheus be bound? corporate global responsibility*. Houndmills, Basingstoke, Hampshire ; New York : Palgrave Macmillan.
- Yin, R. K. (2003). *Case study research: design and methods* (3rd ed). Thousand Oaks, Calif : Sage Publications.
- Yin, R. K. (2009). *Case study research: design and methods* (4th ed). Los Angeles, Calif : Sage Publications.
- Ylijoki, O. et Porras, J. (2016). Perspectives to Definition of Big Data: A Mapping Study and Discussion. *Journal of Innovation Management*, 4(1), 69-91.
- Zikopoulos, P. (2012). *Understanding big data: analytics for enterprise class Hadoop and streaming data*. New York : McGraw-Hill.