HAL
open science

Explorer, visualiser, décider : un paradigme méthodologique pour la production de connaissances à partir des big data

Eglantine Schmitt

To cite this version:

Eglantine Schmitt. Explorer, visualiser, décider : un paradigme méthodologique pour la production de connaissances à partir des big data. Histoire, Philosophie et Sociologie des sciences. Université de technologie de Compiègne, 2018. Français. NNT : . tel-01960545

HAL Id: tel-01960545

https://theses.hal.science/tel-01960545

Submitted on 19 Dec 2018

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

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

Sorbonne Universités – Université de technologie de Compiègne
Laboratoire COSTECH

Explorer, visualiser, décider : un paradigme méthodologique pour la production de connaissances à partir des *big data*

Thèse pour obtenir le grade de Docteure
de l'Université de technologie de Compiègne en
Epistémologie

présentée et soutenue publiquement par
Eglantine Schmitt
le 14 novembre 2018

sous la direction de
Bruno Bachimont

Devant un jury composé de :

- Bruno Bachimont, Enseignant-chercheur HDR, Philosophie des sciences, Sorbonne Universités (directeur de thèse)
- Anouk Barberousse, Professeure des universités, Philosophie des sciences, Sorbonne Universités (rapporteure)
- Emmanuel Didier, Chargé de recherche HDR, Sociologie, ENS Paris (examinateur)
- François-Xavier Guchet, Professeur des universités, Philosophie, Université de technologie de Compiègne (président du jury)
- Sabina Leonelli, Professor, Philosophy and History of Science, University of Exeter, Royaume-Uni (rapporteure)

Explorer, visualiser, décider : un paradigme méthodologique pour la production de connaissances à partir des *big data*

Eglantine Schmitt

Table des matières

Table des matières ... 3

Table des illustrations ... 6

Résumé ... 9

Remerciements ... 10

Introduction ... 12

Première partie. La mythologie de la donnée ... 19

Chapitre I. Les *big data* entre technologie et mythologie ... 20

1. Le phénomène *big data* comme technologie prémoderne de la donnée ... 21

2. Une notion aux mains d'une multiplicité d'acteurs ... 24

3. Le problème logique de la définition ... 29

4. Les *big data* comme mythologie ... 31

5. Les mythes fondateurs ... 33

6. Les composantes du mythe ... 36

Chapitre II. Un imaginaire épistémologiquement chargé ... 40

1. La thèse du renouvellement des acteurs de la production de connaissances ... 44

2. La thèse de la fin du régime explicatif ... 46

3. La thèse de l'empirisme sans modèle ... 48

4. La thèse de l'exhaustivité des données ... 53

5. La thèse de l'abandon de la recherche des causes ... 57

6. La thèse de l'inductivisme ... 62

Conclusion ... 65

Deuxième partie. Traces et calculs dans les sciences computationnelles de la culture ... 66

Chapitre III. De l'empreinte à la donnée du web : une ontologie de la trace numérique ... 67

1. L'épistémologie des sciences de la culture 70
2. La trace dans les sciences indiciaires 72
3. Le log ou journal d'utilisation 76
4. La trace numérique comme log du web 79
5. L'intentionnalité des traces numériques 82
6. Les données comme traces théoriquement chargées 84
Chapitre IV. Vers une continuité épistémique ancrée dans les sciences de la culture 88
1. Le numérique dans les sciences de la culture 90
2. Terrain et corpus : deux modes d'accès au réel 92
3. Les *virtual methods*, une reconstruction de l'accès au terrain 95
4. Les défis de la révolution technique du texte 104
5. Du web comme objet au web comme corpus 114
Conclusion 122
Chapitre V. Le statut du calcul dans les sciences des données 125
1. Rendre les données calculables 126
2. Le champ du calculable 130
3. La pensée computationnelle 134
4. Le calcul dans les sciences de la nature 139
5. De l'analyse exploratoire à la fouille de données 148
6. Le creuset institutionnel et théorique des sciences des données 156
7. La contribution de l'intelligence artificielle 163
Conclusion 169
Troisième partie. Interpréter, visualiser, décider 171
Chapitre VI. Un cas concret d'exploitation de données massives 172
1. Présentation de l'étude de cas 173
2. Expérience client et codification de verbatims 177
3. De la démarche représentative au projet « *big data* » 180
4. La préparation des données 182
5. La conception du plan de code 185

6. La constitution des ressources linguistiques 188
7. L'alimentation du plan de code 192
8. La mise en visibilité de l'activité linguistique 196
9. La mise en histoire et l'interprétation des données 198
Conclusion 203
Chapitre VII. Du récit au système : les formes de restitution de l'analyse de données 205
2.1. Deux configurations épistémiques 206
2.2 La genèse des systèmes 211
Conclusion 220
Chapitre VIII. L'art de la visualisation de données 223
1. La sémiologie graphique 223
2. La cartographie de données 232
3. La représentation comme preuve 237
4. Le langage visuel du design dans la configuration en système 241
5. Matérialisation et interprétation de la recommandation 247
Chapitre IX. Connaissance et décision 250
1. Les *big data*, une épistémologie pragmatiste ? 251
2. Vers une validité graduelle des connaissances 257
3. Décider, connaître et mesurer 261
Conclusion 266
Glossaire conceptuel 270
Bibliographie 274

Table des illustrations

Figure 1. Différentes modalités d'échantillonnage dans les sondages en ligne. Tiré de Frippiat et Marquis 2009 99
Figure 2. Cartographie thématique des digital humanities anglophones, d'après Schmitt (2015).... 109
Figure 3. Cartographie thématique des digital humanities francophones, d'après Schmitt (2015).. 110
Figure 4. Extrait de Exploratory data analysis, John Tukey, 1977 (p. 9) 152
Figure 5. Valeurs et représentations statistiques des ensembles de données du quartet d'Anscombe. Tiré de Tufte, Edward, The Visual Display of Quantitative Information, p 13. 153
Figure 6. Nuages de point des ensembles de données du quartet d'Anscombe. Tiré de Tufte, Edward, The Visual Display of Quantitative Information, p 14. 153
Figure 7. Notes de cours de Robert Tibshirani, professeur de « data sciences biomédicales et statistiques » à l'université de Stanford. 161
Figure 8. Le Perceptron Mark I au Cornell Aeronautical Laboratory 166
Figure 9. Exemple de visualisation d'un score NPS dans un tableau de bord. Le score NPS est une mesure quantitative de la satisfaction client exprimée sur une échelle de -100 à +100. Il est calculé à partir de la note moyenne des clients satisfaits moins celle des clients insatisfaits. 179
Figure 10. Exemple de visualisation de type « treemap » d'un plan de classement. Les codes sont libellés et hiérarchisés dans une organisation à deux niveaux. Sur la visualisation, la surface couverte par chaque zone représente le nombre de verbatims correspondants. 180
Figure 11. Les 20 premières entrées de la gazette Prix utilisée pour les projets dont le client appartient à la grande distribution. On retrouve le score déterminant si l'observable lexical doit être détecté, des redirections, des activateurs et des inhibiteurs. Ainsi, sur la première ligne, l'adjectif « élevé » sera annoté pour la notion de prix s'il est précédé ou suivi du mot « prix » ; cette annotation renverra vers une entité nommée « trop cher ». 189
Figure 12. Capture d'écran du logiciel montrant une vue hiérarchique du plan de code (volet central), le plan de code intermédiaire Problèmes (volet latéral gauche) et la métadonnée Région (volet droit). 197

Figure 13. Exemple de plan factoriel mettant en relation le vocabulaire des réponses à une question ouverte (représenté par des lettres A, B, C, etc.) et les réponses aux questions fermées (sexe, âge et opinion). Tiré de Cibois, 1990. 201
Figure 14. Deux visions de la combinaison de compétences du data scientist idéal, la première par le consultant Drew Conway, et la seconde par Dr Stephan Kolassa, « Data Science Expert ». 216
Figure 15. « Nuage de points » représentant les variables « temps passé sur le site » et « nombre d'amis » d'un utilisateur. La corrélation positive entre les variables (lorsque l'une augmente, l'autre augmente également) est aisément lisible. 225
Figure 16. Graphique tiré du New York Times du 1e février 1976, présenté par Edward Tufte (Ibid.) comme un exemple malhonnête de visualisation de données visant à suggérer que le budget prévu pour 1977 est supérieur à toutes les années précédentes. 229
Figure 17. Le même graphique que la **Figure 16**, redessiné par Tufte suivant ses préceptes. De nombreux éléments ont été supprimés dont les effets d'ombres, les bordures, les annotations, certaines échelles. Les valeurs indiquées ont également été modifiées. Par ailleurs, la population de l'État de New York a augmenté et le dollar a subi une inflation substantielle dans l'intervalle représenté. De ce fait, une mesure différente est utilisée pour montrer l'évolution du budget en prenant en compte ces facteurs : la dépense par habitant en dollars constants. Le ratio consacré aux aides locales, qui exprime une autre idée, a été éliminé ; enfin, une annotation met en évidence que la variable exprimée a diminué de 5% entre 1976 et 1977, contrairement à ce que suggérait le graphique original. 230
Figure 18. Heatmap réalisée avec le logiciel de Proxem, représentant les types de problèmes identifiés sur un ensemble de verbatims clients sur des enseignes de prêt-à-porter, croisés avec l'évolution du volume dans le temps. La taille des cercles représente le volume de verbatims correspondant à chaque combinaison, tandis que l'intensité de la couleur représente un score de surreprésentation. Plusieurs zones attirent l'attention, comme par exemple les problèmes de délai en mars 2015, les problèmes sur la collection entre août et décembre 2015, ou encore l'insatisfaction générale en mars 2016. 233
Figure 19. Le même graphe, représenté deux fois 1) sans ajustement ni spatialisation, tel qu'il apparaît lorsque le fichier est chargé dans le logiciel Gephi. 2) après le travail de « manipulation exercée » consistant à lui appliquer des actions de coloration, spatialisation, filtrage, mise en forme, etc. Le graphe « brut » est illisible tandis que le graphe retravaillé suggère des interprétations. 240
Figure 20. Photo d'un aspirateur Dyson (modèle Dyson DC52 Allergy Pro). 242
Figure 21. Captures d'écran de deux variantes de la page d'accueil de Netflix présentées dans The Netflix Recommender System : Algorithms, Business Value, and Innovation (2015) un article de Carlos Gomez-Uribe et Neil Hunt, tous deux travaillant chez Netflix. La première mise en avant varie (The Others dans la première capture d'écran, The Universe: Collection dans la seconde). La première liste de recommandations est d'un côté une liste de « films à suspense », de l'autre les « top picks » pour l'utilisateur. Les listes de recommandations suivantes sont également différentes. 244

Figure 22. Capture d'écran de l'interface de Netflix, montrant une partie de la liste de « films historiques » recommandés. La première proposition n'est pas un film historique mais un documentaire (Emmanuel Macron, les coulisses d'une victoire) portant de plus sur un sujet contemporain. 248

Résumé

L'enjeu de cette thèse de philosophie des sciences était de répondre à un problème pratique, qui s'est présenté à moi alors que je travaillais comme analyste sur des données massives chez un éditeur de logiciel : comment produire des connaissances valides en manipulant de grandes masses de données que je n'ai pas constituées et qui ne sont pas le fruit d'une méthode scientifique reconnue ?

En m'appuyant sur mon expérience de terrain, je propose un paradigme méthodologique pour la construction, l'exploration et l'interprétation des données massives. Par paradigme méthodologique, on entend un cadre théorique et pratique qui fournit autant de clés pour développer une méthode adaptée aux données et au projet épistémique envisagés. En faisant la part du mythe des big data et des pratiques effectives, je montre comment les données numériques, toujours déjà manipulables, sont construites techniquement et épistémologiquement à partir des traces laissées par les individus et leurs médiations sur les supports informatiques. Cette construction s'appuie sur une logique de constitution qui requiert un cadre interprétatif et une continuité épistémique entre la donnée et les connaissances que l'on cherche à produire. Les sciences de la culture fournissent ainsi un cadre nécessaire, mais pas suffisant, à assurer cette continuité. Le calcul, incarné par les sciences des données et l'intelligence artificielle, actualise et instrumente cette continuité au prix du renoncement à s'envisager comme fin en soi. Ni le cadre théorique, ni le calcul, ne suffisent toutefois à rendre intelligibles les connaissances ainsi produites : c'est la médiation de l'interprétation, du récit et de la conception logicielle (ou design) qui matérialise, donne à voir et contextualise les connaissances ainsi produites. Ces dernières, enfin, ne sont pas légitimées par leur pur caractère véridictionnel, mais par leur capacité à proposer ou faciliter des décisions d'action, bien souvent en entreprise, elles-mêmes productrices de nouvelles traces numériques manipulables.

Remerciements

A mon père, qui n'en croirait pas ses yeux.

Entamer une thèse lorsqu'on se trouve déjà dans une situation professionnelle confortable et prometteuse est une décision insensée. Je remercie chaleureusement Fabienne, Guylain et Marie de toute l'énergie qu'ils ont déployée pour essayer de m'en dissuader. L'assurance avec laquelle j'ai rejeté leurs arguments parfaitement rationnels m'a rendue suffisamment certaine de ma décision pour ne jamais la regretter. Merci une seconde fois à Marie pour son aide irremplaçable dans les débuts de cette thèse, et pour la rigueur intransigeante avec laquelle elle m'a mise sur les rails épistémologiques.

Permettre à une salariée de son entreprise de se lancer dans une thèse est une autre décision insensée. Je remercie vivement François de m'avoir fait confiance et de m'avoir laissée l'imiter dans une discipline qui n'était pas la sienne, peu après sa propre soutenance. J'ai aimé jongler jour après jour entre la philosophie, l'informatique et le marketing en essayant de jouer ce rôle de "pont entre les mondes" qui m'est si cher.

Bruno Bachimont a eu dès le début toute l'admiration intellectuelle que je peux porter à quelqu'un. En lisant sa propre thèse d'épistémologie, il m'est apparu comme une évidence que mon directeur de thèse ne pourrait être que lui. Je lui suis infiniment reconnaissante d'avoir accepté de diriger mes travaux et de m'avoir ouvert la porte du monde de la recherche. Je me réjouis que cette thèse m'ait permis de découvrir qu'il mérite bien mon admiration mais aussi toute ma sympathie. Merci donc d'avoir non seulement défriché la jungle de l'épistémologie du numérique, mais de m'avoir supportée et soutenue jour après jour comme doctorante.

Merci à toutes celles et ceux que j'ai pu croiser à l'UTC, au Costech, et notamment aux équipes EPIN et CRED qui m'ont permis de faire la thèse que j'avais envie de faire. Merci pour votre bienveillance, votre écoute, et l'intérêt que vous avez pu porter à mes travaux. Un merci tout particulier à Cléo d'avoir ranimé en moi la flamme morale.

Anouk Barberousse est aujourd'hui l'une des figures incontournables et tutélaires de la philosophie des sciences française, et l'une des rares à s'intéresser aux enjeux contemporains du numérique. Je lui suis éternellement reconnaissante d'avoir accepté d'être rapporteure de mon travail.

Personne ne s'est aussi bien emparé à ce jour de la question de la donnée du point de vue de l'épistémologie et des STS que Sabina Leonelli, et je suis également très honorée qu'elle ait accepté d'évaluer mon travail. Sabina, I am forever in your debt.

Emmanuel Didier est pour moi l'héritier d'Alain Desrosières dont il a repris le flambeau salutaire. Plus que jamais, il nous faut veiller aux rapports de pouvoir inscrits dans les nombres et je suis très heureuse que ma thèse bénéficie aussi de ce regard.

J'ai trouvé en Xavier Guchet quelqu'un qui n'a pas seulement une connaissance à la fois vaste et précise de la philosophie des techniques, mais aussi l'auditeur idéal, doté d'une bienveillance et d'une capacité d'écoute et de compréhension sans égales. Xavier, c'était un plaisir de te présenter mes travaux.

Merci à Amélie, Benjamin, Benoît, Célya, Fabienne, Fanny, Lucie, Maeva, Marie, Martin, Rémi, Sara, Stéphanie, pour leurs yeux acérés de relectrices et relecteurs.

Merci à toute l'équipe de Proxem pour sa gentillesse et la solide compréhension de l'informatique qu'elle m'a apporté.

Merci à toutes mes amies et amis de m'avoir supportée aussi longtemps. Vous avez toutes et tous très bien fait semblant de vous intéresser à mon sujet. Vous n'en étiez pas, pour la plupart, à votre première thésarde, mais vous avez accepté mes silences comme mes obsessions et ça aurait été beaucoup plus difficile sans vous.

Merci à Alexandra Elbakyan sans laquelle la bibliographie de cette thèse serait significativement moins fournie.

Merci à Twitter et Netflix qui ont été des respirations envahissantes mais indispensables.

Enfin, je remercie Thomas, partenaire infaillible de toutes mes aventures, qui ne sait même pas combien il m'a été nécessaire, du plus fort que je peux. Tu ne m'as pas seulement soutenue jour après jour, mais aidée, éclairée, rassurée, conseillée. Cette thèse aurait sans nul doute été, à tous points de vue, beaucoup moins bien sans toi.

Introduction

> Dans la littérature philosophique, on trouve une foison d'explications compliquées des théories causales de la perception, mais elles sont curieusement éloignées de la vie réelle. Nous avons des descriptions fantastiques de chaînes causales aberrantes qui, à la manière de Gettier, remettent en question telle ou telle analyse conceptuelle. Mais le microscopiste moderne connaît des tours bien plus étonnants que le plus imaginatif des spécialistes de philosophie de la perception. Ce qui nous manque en philosophie, c'est d'être plus conscients de ces vérités qui sont plus étranges que des fictions. II serait bon que nous ayons quelques lumières sur ces extraordinaires systèmes physiques « dont le pouvoir grossissant nous permet aujourd'hui de voir plus que tout ce que l'on a jamais pu voir auparavant dans le monde ». (Ian Hacking, « Est-ce qu'on voit à travers un microscope ? », 1981)

Chaque innovation dans les technologies de la connaissance bouleverse notre rapport au réel, augmentant notre perception, notre mémoire et nos capacités de raisonnement. Les instruments de mesure scientifiques dédiés à l'observation dévoilent de nouveaux pans du réel, tandis que les outils dédiés à la manipulation nous donnent la capacité à intervenir dans ce qui n'est dès lors plus la nature immaculée, mais un système constitué de ce que nous y avons trouvé et de ce que nous y avons apporté. Le télescope nous a donné accès au lointain, le microscope à l'infiniment petit, la radiographie à l'intérieur de la matière. Plus près de nous, l'avènement du numérique a réinventé la façon dont nous consignons et partageons nos connaissances. Il est un nouveau support pour l'agir et le savoir, en même temps qu'un nouvel outil pour manipuler et constituer ce savoir. La multiplication des traces que nous laissons de nous-mêmes sur ces supports numériques nous donne désormais un accès inédit à notre propre culture.

Chaque innovation dans les technologies de la connaissance appelle une nouvelle épistémologie, un nouveau regard raisonné sur les objets que nous souhaitons connaître. Tout en s'appuyant elles-mêmes sur des connaissances, ces technologies augmentent la pensée productrice de connaissances, et ses capacités de mémoire, d'apprentissage et de manipulation. La technique ne fait pas qu'instrumenter l'esprit scientifique, mais le pousse hors de ses retranchements, vers de nouvelles théories du réel et de nouvelles méthodes pour l'appréhender. Ces approches nouvelles, hésitantes et

bancales, construisent néanmoins des ponts entre ce que nous pouvons voir et manipuler, et notre exigence de rationalité. Comme l'écrit Karl Popper (1985),

> La raison procède par essais et erreurs. Nous forgeons des mythes et des théories, que nous essayons ensuite : nous tentons de voir jusqu'où ceux-ci nous conduisent.

À ce titre, les approches nouvelles suscitées par les innovations dans les technologies de la connaissance sont forcément insatisfaisantes, au prisme de nos normes habituelles comme du fait de leur caractère naissant. Elles sont toujours incomplètes, insuffisantes, inacceptables. Elles seront critiquées, amendées, revisitées, reprises à la racine. Néanmoins, sans l'imperfection de ces essais pionniers, il n'y a rien d'autre à parfaire que la déconstruction de ce qui aurait pu être fait.

La multiplication des traces que nous laissons de nous-mêmes sur les supports numériques ne fait pas exception à ces constats. Désignée presque indifféremment sous les termes de *big data*, de *data sciences*, d'*algorithmes*, d'*intelligence artificielle*, l'étude raisonnée et techniquement instrumentée de ces traces émerge avec son cortège de « mythes et de théories », comme le dit Popper, que nous formulons en chemin. Comme à l'aube du *logos* platonicien, la frontière entre le mythe et la science est encore fragile, et il faut un regard perçant pour les distinguer. Le mythe raconte une histoire plus plaisante et plus facile à comprendre que les tâtonnements pleins de détails techniques des premières réalisations, et se propage ainsi plus vite et plus fort. L'expérimentateur navigue à vue, autant à partir de ce qu'il sait qu'à partir de ce qu'il aimerait savoir, entremêlant les deux. Il est lui-même le héros du mythe qu'on lui raconte, et qu'il se raconte pour s'orienter. Bien qu'elle s'en inspire, il n'y a, comme on le verra, rien dans l'étude des traces numériques qui satisfasse aux critères des sciences contemporaines quelles qu'elles soient, tout en ayant un mode d'émergence fondamentalement similaire.

Pour comprendre ce qui se joue avec la multiplication des traces numériques, il nous faut prêter l'oreille au mythe agréable autant qu'aux détails techniques, et les prendre au sérieux tous les deux. Pour rendre raison de ces nouvelles technologies de la connaissance, il nous faut mobiliser une philosophie des sciences bienveillante et attentive aux détails. Il nous faut opter pour une attitude simultanément descriptive et normative, car décrire une première fois les choses, c'est les nommer, c'est-à-dire poser les termes dans lesquels elles peuvent être appréhendées. Faire l'épistémologie des *big data*, c'est construire l'appareillage théorique et la posture conceptuelle requise pour comprendre et faire l'étude de ces grandes masses de données. C'est formuler un premier **paradigme méthodologique**, suffisamment général pour s'appliquer à tout projet d'étude de ce type, et suffisamment spécifique pour orienter déjà les ajustements nécessaires à la situation effective d'un projet.

Cette approche simultanément descriptive – presque historicisante – et normative – au sens où sa contribution est un paradigme prescripteur de méthode – nous permet d'échapper à une autre

normativité, moins fructueuse : celle de la déconstruction. Comme nous le verrons, dire d'un phénomène qu'il est une construction sociale, c'est l'enfermer et le condamner, comme s'il n'avait plus rien à dire au-delà de son statut d'artefact culturel. Le réduire à un objet sociologisant, support de jeux de pouvoir entre acteurs, c'est le ramener à une pure extériorité. Pour nous éloigner de cet écueil, nous faisons le choix d'une approche philosophique qui intègre les propriétés internes et externes de l'objet, dévoile la complexité du système de croyances qu'il constitue, plutôt que de le ramener dans les frontières d'un objet simple, sans pli, qui ne supporte qu'un angle d'analyse.

Par ce parti pris, c'est une certaine conception de la philosophie des sciences qui est mobilisée. Une certaine conception de la philosophie, tout d'abord, consciente des critiques qui lui furent adressées et de sa prédilection pour les « explications compliquées […] curieusement éloignées de la vie réelle » selon la formulation de Ian Hacking, lui-même philosophe, et cité en exergue de cette introduction. La philosophie que nous pratiquons se veut sans cesse nourrie du réel par l'intermédiaire de ce qu'ont à en dire les sciences humaines et sociales, et par l'expérience de première main quant à son objet dont bénéficie, comme on le verra, l'auteure de ces lignes. C'est une philosophie qui se veut définie par l'objet qu'elle se donne et les développements qu'il lui impose, plus que par une tradition philosophique particulière, une école de pensée spécifique. Nous aurons l'obsession de ce qui s'est réellement produit, de ce qui est observable, des conditions matérielles, pratiques et empiriques d'émergence de notre objet. Notre ancrage se situe dans la charnière entre science et technologie articulée par les STS *(science and technology studies).* Il s'agira autant d'une philosophie des sciences qu'une philosophie des techniques, adoptant une approche conceptuelle sans relever de la philosophie analytique, aussi bien qu'une approche historicisante, sans relever de l'histoire des sciences.

Nous mobilisons également une certaine conception de la science (ou *des* sciences), suivant laquelle elle n'est pas réduite à un chapelet de disciplines soumises à l'impérialisme scientifique (Mäki 2013) de la physique, mais intègre toute étude simultanément empiricisante et raisonnée d'un objet, qu'elle relève de la nature ou de la culture, qu'elle procède d'institutions scientifiques bien établies ou d'amateurs débutants. Nous nous efforcerons également d'éviter à cette conception l'écueil de la naïveté selon laquelle les pratiques épistémiques sont pures et désintéressées, au service d'un dévoilement de la vérité comme correspondance au réel. Nous considérerons ainsi que les acteurs de ces pratiques, tout en s'inscrivant dans un certain réalisme scientifique plus ou moins sophistiqué, agissent également selon d'autres normes épistémologiques et extra-épistémologiques.

Notre objet n'est donc pas la connaissance pure et désintéressée qui pourrait se dégager de l'amoncellement des traces de nos activités ; nous envisageons la question des *big data* comme un système, avec des composantes épistémiques, mais aussi pratiques, économiques et sémiotiques. L'innovation dans les technologies de la connaissance s'accompagne notamment d'enjeux économiques qui ont de sérieuses conséquences sur la production effective de savoirs. Ces technologies ont un coût, et se vendent plus qu'elles ne se donnent. Ceux qui y ont accès ne sont pas

forcément ceux qui en bénéficieraient le plus ou le mieux. Sans nous livrer à une cartographie détaillée des agents et flux financiers concernés par ce système, nous aurons toujours à l'esprit ces conditions pratiques, non seulement techniques mais aussi économiques, notamment en ce qu'ils apportent comme facteurs d'explication des modalités de production de connaissances.

Comme tout bon objet philosophique, le phénomène *big data* est vaste, riche et complexe. Comme peu d'objets philosophiques en revanche, peu de choses ont été dites sur lui, qui valent la peine d'être retenues : c'est peu ou prou une terre vierge pour l'épistémologie. Notre ambition n'est donc pas de l'épuiser, de le conceptualiser dans sa totalité. Inscrit dans une temporalité, il n'est plus ce qu'il était au début de nos recherches, et n'est sans doute pas près de se stabiliser. Plus modestement, notre ambition est d'apporter, comme les défricheurs, des premières clés de lecture pour prendre la mesure de l'objet, de sa complexité, des différents angles sous lesquels on peut l'envisager, et permettre à d'autres de s'épargner des explorations spéculatives ou stériles. Ces clés de lecture sont autant conçues comme des moyens de comprendre, que des remèdes au risque de mal comprendre un sujet visé par beaucoup de discours superficiels, émotionnellement ou axiologiquement chargés. Nous souhaitons, en particulier, clarifier ce que ces technologies rendent effectivement possible et ce qui relève encore aujourd'hui de la science-fiction, genre qui alimente largement les mythes des *big data*, et dont un auteur, Arthur C. Clarke, a souligné dans un autre contexte, que toute technologie suffisamment avancée paraît indiscernable de la magie. Si au terme de ce travail, le lecteur a clairement à l'esprit ce qui appartient aux possibilités technologiques contemporaines « plus étranges que la fiction » et ce qui relève de la magie, nous aurons déjà accompli quelque chose.

Les clés de lecture que nous proposons s'articulent autour du problème d'élucider les conditions de possibilité et les modalités de validation des connaissances issues du traitement de *big data*. Il s'agit ainsi, en étudiant des pratiques effectives, de comprendre comment de telles connaissances sont produites et acceptées. Cette question ne peut toutefois être résolue sans prendre en considération le cadre sociologique et discursif dans lequel s'inscrivent les pratiques des acteurs : nous procédons ainsi préalablement à une analyse des *big data* comme discours dont la puissance rhétorique n'est pas sans effet sur les pratiques effectives. L'élucidation de ces mécanismes rhétoriques nous permet d'arriver à notre seconde et principale question, celle des modalités et frontières des connaissances effectivement produites. Nous proposons ainsi une *critique* de ces connaissances, au sens kantien de détermination de leurs frontières et de leur domaine de validité. Ce travail critique permettra non seulement de rendre compte des différents aspects de la production de connaissances à partir de traces numériques, mais d'organiser cette contribution en proposant formellement un **paradigme méthodologique** fournissant un cadre de travail pour des pratiques ultérieures. Dans ce but, nous présenterons trois contributions complémentaires :

- une **épistémologie de ces traces** abondantes qui sont toujours déjà appelées des données, et dont la valeur épistémique dépend presque autant de la signification qui leur est attribuée que de la mythologie qui leur est associée ;
- un examen des **techniques computationnelles et méthodes d'analyse**, constituées en savoir-faire plus qu'en science, qui visent à apporter à ces traces une intelligibilité, et qui constituent les conditions matérielles de possibilité de cette intelligibilité ;
- une investigation, ancrée dans le réel par un exemple détaillé, des modalités de représentation et d'**interprétation** des données, ainsi que des normes de **validation** qui régissent ces herméneutiques alimentées par les traitements computationnels.

Pour ce faire, nous naviguerons, toujours guidés par la philosophie des sciences, entre une analyse des discours, des travaux scientifiques et des objets numériques directement observables. Nous emprunterons ainsi aux méthodes de l'histoire et de la sociologie des sciences, de la bibliométrie, de la sémiologie et de l'ethnographie, tout en gardant comme cadre celui de l'approche philosophique qui est la nôtre.

En particulier, la première partie entend mettre à distance les discours mythologiques relatifs aux *big data*, tout en prenant la mesure du rôle de ces discours, abondants et décisifs, sur les pratiques effectives. Nous étudions ainsi les *big data* comme un phénomène discursif ancré dans un contexte culturel, qui se présente notamment comme une remise en cause des régimes de scientificité traditionnels dans les sciences de la nature. Le premier chapitre mobilise ainsi l'analyse de discours pour déterminer le statut de ces discours et faire émerger, à partir des mythes des *big data*, des arguments d'ordre épistémologique que nous pourrons ensuite (chapitre II) mettre en perspective, voire critique (au sens kantien) ou invalider. Ces arguments sont ensuite mis en regard des pratiques, qui s'inscrivent dans l'épistémologie des sciences de la culture plus que dans celles de la nature.

La seconde partie propose la notion hypothétique de sciences computationnelles de la culture pour conceptualiser des pratiques rationnelles de traitement de données massives fondées sur une continuité épistémique entre la donnée, les outils et méthodes pour la manipuler, et le cadre conceptuel et théorique dans lequel s'inscrivent ces traitements. Nous proposons ainsi les notions de trace et de paradigme indiciaire (chapitre III) pour caractériser les données étudiées et la rupture toujours déjà présente qu'elles entretiennent avec l'objet qu'elles sont présumées représenter. La trace numérique permet de produire des connaissances nomologiques, portant sur le général, comme des connaissances idiographiques, mettant en évidence le particulier. Cette production de connaissances requiert une continuité épistémique entre les données, les méthodes et le cadre conceptuel, afin de faire émerger des sciences computationnelles de la culture. Cette continuité est à ce jour introuvable dans les sciences de la culture historiquement attestées qui mobilisent des traces numériques (chapitre IV). Elles proposent néanmoins plusieurs façons d'ancrer la donnée dans une approche disciplinaire, en particulier à partir de la notion de corpus. Malgré cette notion, il y a toujours déjà une rupture

épistémique entre la donnée et ses outils d'analyse, désormais envisagés comme des composantes séparées de notre paradigme méthodologique. Les sciences des données fournissent ces outils d'analyse (chapitre V), mais sans les articuler avec un cadre théorique spécifique : contrairement aux approches computationnelles dans les sciences de la nature, les sciences des données ne sont pas reliées à une théorie scientifique à travers la notion de modèle, mais fonctionnent comme une boîte à outils a-théorique, mobilisée avec une approche exploratoire. Nécessaires mais pas suffisantes pour produire des connaissances à partir de traces numériques, elles appellent en creux une composante supplémentaire.

La troisième partie examine les conditions d'intelligibilité et de validité des connaissances produites à partir du calcul sur les données, à travers l'interprétation, la narration et la visualisation. La mise en évidence de la composante supplémentaire requise est faite à travers une étude de cas (chapitre VI) qui remobilise la trace numérique et les sciences des données, mais fait également apparaître des formes de restitution spécifiques, une multitude de culture épistémiques, et un savoir-faire interprétatif. Deux formes de restitution sont compatibles avec les sciences computationnelles de la culture (chapitre VII) : le récit, qui rend raison de l'investigation faite sur et avec les traces numériques, et le système, qui suggère diverses investigations sous forme logicielle, et une figure supplémentaire, celle de l'utilisateur, capable d'interagir avec le système. Dans ces deux configurations, la visualisation de données est à la fois un outil heuristique pour l'exploration des traces, un langage capable de les rendre intelligible, et un support de preuve pour les connaissances produites (chapitre VIII) ; dans les configurations en système, le design complète ce langage visuel avec des possibilités d'interactions et des pistes d'interprétation matérialisées dans une interface.

Enfin, quelle que soit leur forme de restitution, les connaissances produites suivant le paradigme méthodologique ainsi construit s'inscrivent dans un régime de validité d'ordre instrumentaliste ou pragmatiste, qui fait de l'actionnabilité plus que de la vérité le critère de validation des résultats des analyses de traces numériques (chapitre IX) : ces résultats, c'est-à-dire les connaissances ainsi produites, suggèrent et légitiment des décisions qui leur confèrent en retour une valeur épistémique définitive.

La double approche descriptive et normative ainsi proposée nous permet de construire un paradigme méthodologique pour l'analyse des traces numériques articulé autour des notions de corpus, d'exploration, de visualisation et de décision : ce paradigme rend compte des pratiques existantes à l'œuvre dans une multitude de contextes institutionnels, tout en ouvrant la voie à des pratiques futures qui sauront adopter ce cadre de travail.

Enfin, puisque nous nous sommes donné comme obsession les conditions pratiques d'émergence des connaissances, il nous faut préciser pour finir celles qui furent celles de cette thèse : une double activité de doctorante à l'Université de technologie de Compiègne – lieu privilégié pour observer l'articulation entre technique et connaissance – et de salariée d'un éditeur de logiciel spécialisé dans les *big data*, du nom de Proxem. Dans cette entreprise, nous avons mené une recherche-action qui conjuguait un travail de recherche universitaire et une activité pratique d'analyse de données textuelles massives régie par la nécessité d'en dégager une intelligibilité effective. Plusieurs missions nous ont été confiées, dont celles de réaliser des études fondées sur le traitement de *big data*, dans les domaines du marketing et des ressources humaines, et d'analyser et comprendre les besoins et usages du logiciel développé par l'entreprise. A travers ces différentes missions, et par immersion dans le *milieu* de cette entreprise, nous avons pu non seulement observer de près, mais aussi faire l'expérience concrète, des pratiques contemporaines émergentes formées par la multiplication des traces que nous laissons de nous-même.

Première partie.

La mythologie de la donnée

Chapitre I.
Les *big data* entre technologie et mythologie

L'apparition de nouvelles masses de données dans les sources de la connaissance pose le problème concret et pratique de déterminer comment de telles données sont identifiées, manipulées et comprises dans le cadre d'une activité épistémique. C'est le problème que nous nous sommes donné dans le cadre de la présente recherche-action. Dès lors qu'il s'agit d'analyser et de comprendre ces mécanismes, une solution naturelle est d'écouter et de prendre au sérieux ce que les praticiens ont à en dire. Cette solution d'apparence logique est toutefois mise en difficulté dès lors que les nouvelles pratiques s'inscrivent dans une économie de la promesse où discours et pratiques s'influencent mutuellement, et où divers régimes de croyances se retrouvent mêlés. La difficulté d'un examen philosophique des *big data* est ainsi qu'elles ne sont pas seulement des objets techniques constitués, dont on pourrait, à la manière de Simondon, examiner le mode d'existence ou la concrétisation. Elles forment un phénomène culturel qui intègre des objets concrets mais aussi un ensemble de discours et de croyances quant aux objets eux-mêmes et à ce qu'ils pourraient être. Ces discours font partie intégrante du phénomène et ont non seulement une existence textuelle et médiatique, mais aussi une influence sur les pratiques concrètes qui constituent notre objet : on ne peut donc étudier ces pratiques si l'on fait l'impasse sur ce qui les motive, les influence, ou prélude à leur constitution. Pour le dire autrement, les discours sur les *big data* ont une puissance rhétorique telle que les praticiens des *big data* y sont exposés, y adhèrent et les véhiculent, et que leurs pratiques sont au moins partiellement motivées ou déterminées par ces croyances.

Ils sont aussi pour ces praticiens un moyen de formaliser des hypothèses et de les discuter, de verbaliser des façons de faire émergentes, d'apporter une forme de rationalité à leurs pratiques grâce à la médiation du langage. Parmi les auteurs de ces discours, on trouve des journalistes, mais aussi des experts, des praticiens, des chercheurs, c'est-à-dire des acteurs directement impliqués dans les pratiques des *big data*. Ces discours permettent à ces acteurs d'expliciter l'horizon d'attente et les concepts au travail dans leurs méthodes : les pratiques influencent les discours, les nourrissent, et en fournissent la substance. C'est enfin la principale *source* dont nous disposons pour examiner un phénomène contemporain largement extrascientifique, dont les formes de restitution ne sont pas

exclusivement celles de la science classique (communications, publications, journaux de laboratoire, etc.) mais aussi une multitude de supports allant de l'article de presse au livre blanc en passant par la tribune, le blog, le tweet, les notes de travail, le logiciel, ainsi que des communications non écrites sur lesquelles nous reviendrons. Cette source première, mais non définitive, va nous occuper dans le présent chapitre et celui qui suit.

Discours et pratiques forment un système, constitutif du phénomène *big data*, dont l'étude doit intégrer ces deux dimensions. L'étude de ces discours apparaît nécessaire en premier abord pour analyser les pratiques qui constituent notre objet, dans la mesure où ces discours auront pour nous une fonction d'explication au moins partielle desdites pratiques.

Nous entendons montrer que ces discours ne sont pas à envisager selon un statut véridictionnel. Cela ne veut pas dire que leur contenu est nécessairement injustifié, ou qu'ils ne mettent pas en lumière des aspects pertinents du phénomène *big data*, mais que, dans notre optique de compréhension des pratiques, nous ne nous prononçons pas sur leur valeur de vérité, que nous la mettons en suspens. Dans cette perspective, nous suggérons, après avoir clarifié le sens que nous donnons à ces termes, qu'ils sont tout aussi bien une *technologie* (dans son sens premier) qu'une *mythologie* (au sens barthésien). Après avoir ainsi mis en évidence son caractère idéologique, nous montrons que la mythologie des *big data* intègre néanmoins un certain nombre de croyances d'ordre épistémologique, des concepts structurants relatifs au statut de la donnée et de son traitement, qui influencent les praticiens malgré leur non véridicité, et constituent donc une clé d'entrée incontournable dans la compréhension des pratiques. Dans le deuxième chapitre, nous examinons ces thèses épistémologiques et montrons en quoi elles sont discutables, réfutables (notamment *via* les publications des praticiens), ou vérifiables logiquement ou au travers d'exemples. Nous verrons ainsi ce que nous pouvons conserver de ces thèses pour caractériser l'objet et construire effectivement des connaissances à partir des données et des traitements à l'œuvre dans les *big data*.

1. Le phénomène *big data* comme technologie prémoderne de la donnée

Dans son sens premier, la technologie désigne « les tentatives pour mettre en forme les savoir-faire artisanaux » (Carnino 2010). Historiquement influencée par la tradition savante de l'encyclopédisme des Lumières, cette conception de la technologie (antérieure à la science moderne qui émerge au XIXe siècle) en fait une science appliquée, voire autonome. En effet, comme le souligne Guillaume Carnino,

> Savoir vraiment, ce n'est pas encore nécessairement connaître les causes scientifiques des phénomènes naturels et productifs, mais c'est cataloguer les différentes manières de faire, matériellement sédimentées dans les objets et incorporées (et souvent théorisées) par les artisans.

La technologie dans ce sens prémoderne désigne donc un savoir fondé sur la description rationnelle des pratiques. C'est une *épistémè* qui s'oppose à la *doxa*. Les savoirs qu'elle désigne apparaissent comme préscientifiques ou infra-scientifiques si on les évalue au prisme de la science moderne, qui s'appuie sur des pratiques scientifiques plutôt que des savoir-faire artisanaux. Il s'agit néanmoins de discours rationnels et méthodiques, de discours *savants* sur des techniques auxquelles on confère une certaine dignité intellectuelle, un caractère différent de celui de l'opinion confuse et du préjugé.

La raison pour laquelle il est donc intéressant d'évoquer ici cette acception aujourd'hui obsolète du mot *technologie* est que les discours relatifs aux *big data* correspondent assez bien, comme on va le voir, à des « tentatives pour mettre en forme les savoir-faire artisanaux », bien qu'ils ne s'y réduisent pas. Ces discours sont de natures variées, mais ils ont en effet en commun de proposer ou propager une légitimation de pratiques infra-scientifiques de traitements de données massives. La mise en forme de ces savoir-faire se fait ainsi sur un mode qui vise à les promouvoir autant qu'à les expliciter. Les raisons de cette volonté de légitimation sont connues : elle s'explique par les motivations des auteurs de ces discours. Ils ne sont en effet pas anodins en termes de charge axiologique puisqu'il s'agit souvent de sociétés de conseil, d'intégrateurs et d'éditeurs de logiciel qui commercialisent notamment ou exclusivement des produits et des offres fondés sur le traitement de données massives et ont donc une motivation économique à légitimer les pratiques des *big data* en général. Ces discours ne proviennent que rarement des institutions scientifiques, ce en quoi ils rappellent encore une fois la notion prémoderne de technologie. Ils sont, en effet, des discours sur les techniques et ne relèvent donc pas de la philosophie naturelle, l'ancêtre prémoderne des sciences contemporaines. Du fait de leur caractère émergent et artisanal, les pratiques *big data* ne satisfont pas les critères d'acceptabilité de la science contemporaine, mais leurs discours d'accompagnement y font cependant souvent référence : cela explique par exemple pourquoi elles sont parfois présentées comme un nouveau paradigme scientifique au sens kuhnien. Historiquement, elles viennent après le développement de la science moderne, et ne forment pas des discours complètement autonomes, ni en faveur d'un retour au préscientifique. Du fait de ce rapport spécifique à la science, il n'est pas étonnant que les auteurs de ces discours ne soient pas exclusivement des savants, des chercheurs, mais également des médias, des amateurs, des industriels, produisant une multitude de points de vue sur un objet.

En anglais, l'expression *big data* désigne littéralement d'importants volumes de données. De nombreux discours mettent en effet en avant le caractère massif des traitements *big data* comme un trait essentiel. Néanmoins, dès le départ, il ne s'agit pas seulement d'une question de quantité de données. Plusieurs sources universitaires (Taylor, Schroeder et Meyer 2014; Lagoze 2014; R. Kitchin et McArdle 2016; Rieder et Simon 2017) et de nombreuses sources en ligne, s'accordent à dire que ce point de départ est la formule imaginée par Doug Laney (2001), alors consultant dans le cabinet de conseil Gartner (anciennement META group), dans un document consacré à l'impact du développement du commerce électronique sur la gestion des données internes des entreprises.

L'expertise de Gartner porte sur les évolutions techniques développées par les entreprises, en particulier dans le secteur informatique. Historiquement, ce n'est pas la première trace écrite de l'utilisation de l'expression, mais la définition de Doug Laney est présentée comme fondatrice de son usage actuel, en association avec une caractérisation mnémotechnique triple autour des « 3V », volume, variété et vélocité.

Néanmoins, si l'on peut considérer l'expression « *big data* » comme un néologisme, ce n'est pas la première fois dans l'histoire des sciences que l'on rencontre des discours s'inquiétant ou s'émerveillant de la quantité ou de la complexité des données disponibles. À titre de comparaison, Hummon et Fararo (1995) décrivaient déjà, il y a 20 ans, des jeux de données présentant plusieurs de ces caractéristiques, et avec un enthousiasme qui n'est pas sans rappeler certains discours contemporains :

> We can usefully employ far greater quantities of data, and thereby tackle data analyses beyond the scope of practicality even a few years ago. For example, desktop machines can analyze all annual surveys in the General Social Survey, from 1972 to the present; the GSS database is about 60 megabytes of data. Perhaps even more important than quantities of data is the ability to manage and use data stored in complex data structures. Modern relational database technology supports almost any data structure sociologists can conceive, e.g. complex kinship structures, or the hypergraph structures of contemporary network analysis (Date, 1987).

Bruno Strasser (2012) suggère pour sa part que ce sentiment d'être submergé par les données se rencontre déjà bien avant, chez les naturalistes du XIXe siècle, et même tout au long de la Renaissance : l'histoire naturelle est une affaire de collecte et d'accumulation de données qui apporte ses problèmes de stockage, d'organisation et d'intelligibilité. En apparence, la révolution apportée par l'augmentation du volume et de la complexité des données est non seulement déjà là depuis quelques siècles, mais elle relève davantage de la routine que de la révolution. Par ailleurs, les caractérisations des *big data* ne désignent pas seulement les données elles-mêmes, mais aussi, comme on va le voir, des capacités de traitement algorithmique, ou encore tout simplement un phénomène global, informe, d'enregistrement et de surveillance des comportements humains.

La définition du Gartner est populaire mais elle est loin de constituer un consensus autour d'éléments nécessaires et suffisants ; il ne s'agit d'ailleurs pas à proprement parler d'une définition mais de critères d'identification. Depuis plus de 15 ans, elle s'enrichit de nouveaux critères et en perd d'autres chemin faisant. Kitchin et McArdle (2016) ont ainsi relevé une multitude de nouveaux traits tels que l'exhaustivité, la finesse de la granularité, le caractère relationnel, l'extensibilité, la capacité à passer à l'échelle, la véracité, la valeur, la variabilité, etc. En pratique, aucun jeu de données ne présente tous ces traits, et presque tous les jeux de données historiquement constitués en possèdent au moins un : nous n'avons donc pas là des critères nécessaires et suffisants qui permettraient d'identifier des *big data*, mais des indices, pour un observateur attentif, tendant à suggérer ou non que les données dont

on dispose relèvent des *big data*. La question du volume, motivée par la traduction de l'expression littérale « *big data* » ne semble pas suffisante à décrire l'objet.

Si l'on souhaite rendre compte des « tentatives pour mettre en forme les savoir-faire artisanaux » des *big data*, il nous faut prolonger notre analyse de la seule question du volume à des problématiques plus larges, c'est-à-dire mobiliser des discours qui soulèvent d'autres questions, d'autres point de vue, d'autres objets, et ce vraisemblablement en prenant en compte la richesse de la provenance de ces discours.

2. Une notion aux mains d'une multiplicité d'acteurs

La richesse des discours relatifs aux *big data* renvoie à une multiplicité d'acteurs, qui est à prendre au sérieux : il ne faut pas la voir comme une complexité au travers de laquelle le philosophe devrait se repérer et trancher, mais comme une caractéristique de ces discours. Lorsqu'on compare en effet les assertions des uns et des autres, on constate qu'il n'y a pas de point de vue qui serait propre à un groupe d'acteurs, mais que toutes sortes de points de vue se retrouvent chez toutes sortes d'acteurs. En particulier, les praticiens n'ont pas, comme on va le voir, une position privilégiée dans cette écologie : il n'y a pas d'un côté, le point de vue de ceux qui ont une expérience pratique de l'analyse de données, et de l'autre, ceux qui n'en ont pas. Les *big data* comme technologie relèvent d'une économie de la promesse qui n'est pas un effort coordonné de ses promoteurs, mais un réseau de positions et de points de vue sur l'objet qui constituent précisément sa complexité.

Nous allons développer ce point en analysant plus précisément la teneur des discours sur les *big data*. Dans ce but, nous explorons les valeurs d'usage de l'expression plutôt que de définir normativement une valeur de référence et de la soumettre à une analyse conceptuelle. L'étude d'exemples d'emplois issus de textes scientifiques ou médiatiques permet ainsi d'établir la variabilité des propositions relatives aux *big data*. À l'issue de cette investigation, des familles de définitions apparaissent et peuvent être articulées d'une façon qui permettra de se livrer ensuite à une analyse critique. On pourra ainsi établir le régime discursif de ces définitions et leur articulation avec les pratiques concrètes d'analyse de données et le contexte sociohistorique de ces discours.

Pour ce faire, notre méthode est empirique : nous recensons et décrivons des usages concrets de l'expression « *big data* » au sein de différents discours, de nature scientifique ou médiatique. Le repère de cette analyse était la présence de l'expression lexicale « *big data* », quel que soit son contexte d'utilisation et le régime d'énonciation du discours qui le contient. Se limiter à l'expression littérale elle-même, plutôt qu'à des synonymes ou des reformulations, peut sembler restrictif, mais cette approche nous permet de nous assurer que nous ne posons aucune hypothèse préalable, à ce stade, quant à la teneur des affirmations des différents acteurs. Dans une perspective prudente, d'approche linguistique, tous les discours qui emploient l'expression « *big data* » s'inscrivent sans ambiguïté dans

notre problématique : affirmer qu'un discours puisse employer l'expression de manière impropre nous ferait basculer d'un point de vue descriptif à un point de vue normatif. Si l'expression est employée dans un discours ou un document, il faut considérer que ce support entre *de facto* dans notre corpus.

Ainsi, pour rendre compte des contextes d'utilisation de l'expression, nous avons ainsi analysé la façon dont différentes familles d'acteurs, professionnels, scientifiques et médiatiques, définissent l'expression « *big data* ». Par définition, on entend une formulation de type « les *big data* sont ___ », « les *big data* représentent ___ » ou d'autres paraphrases de ce type. Le corpus, détaillé en annexe, a été constitué comme suit.

- Pour le monde professionnel, nous avons puisé dans un article du blog « datascience » de la *School of Information* de Berkeley. En vue de la publication de cet article, intitulé « What is *big data*? »[1], quarante-trois professionnels d'horizons variés et revendiquant une expertise technologique ont été interrogés et ont proposé leur définition des *big data* ;
- Pour le monde scientifique, une bibliographie non exhaustive, mais délibérément disparate en termes de discipline et de culture scientifique des auteurs, a été constituée par veille personnelle, et par rebond à partir d'articles souvent cités, comme celui de boyd et Crawford (2011) ;
- Enfin, pour le monde médiatique, une vingtaine d'articles journalistiques a été identifiée *via* une veille au fil de l'eau entre janvier 2012 et mars 2014, en privilégiant les contenus plus originaux qui ne se contentaient pas de relayer des points de vue déjà rencontrés.

Le corpus représente ainsi un total de 107 définitions en français (12%) ou en anglais (88%), réparties 43 définitions de professionnels, 42 articles scientifiques et 22 articles de presse. Dans l'ensemble, quatre familles de définitions, que nous allons détailler ci-dessous, se dégagent :

- autour des données elles-mêmes ;
- autour des outils et méthodes d'analyse ;
- autour des usages et fonctions ;
- autour de leur rôle culturel.

Elles se déclinent de la façon suivante :

- **Les *big data* désignent les données elles-mêmes**. L'expression « *big data* » désigne le plus souvent, à l'exception d'autre chose, les données elles-mêmes, qui sont caractérisées de diverses façons. Massives, elles ne sont pas tant une quantité précise que quelque chose qui excède, déborde, dépasse. Elles sont « vastes », « massives » en nombre trop important pour les outils de stockage traditionnels. Elles mettent en difficulté celui qui les administre, celui

[1] https://datascience.berkeley.edu/what-is-big-data/ (consulté le 10 novembre 2017)

qui les analyse, ou encore celui qui essaie de comprendre leur signification et ce qu'elles représentent pour notre époque. Cette profusion n'est pas toujours chiffrée, mais elle l'est parfois abruptement. Ainsi, pour le fondateur de la startup Datahero, les *big data* commencent à 10 To de données. Ailleurs, le seuil est atteint lorsque l'on excède la capacité de stockage d'un ordinateur unique (d'après l'Association for Computing Machinery), ou encore lorsque la manipulation du jeu de données met en difficulté le fonctionnement de Microsoft Excel (pour Steven Weber professeur de sciences politiques à la UC Berkeley School of Information) : volumes conséquents donc, mais pas toujours de la même façon. Après la question de la masse et de l'excès pour ainsi dire « bruts », sans conceptualisation, vient tout ou partie de la définition de Doug Laney. Plusieurs variations autour de cette définition mobilisent des notions plus ou moins synonymes comme la variabilité, l'hétérogénéité, la simultanéité, la dimensionnalité, la forte granularité, etc., recensées et complétées notamment par Rob Kitchin and Gavin McArdle (2016). On rencontre par ailleurs un certain nombre de caractérisations qui font des « *big data* » un genre spécifique de données. D'autres extraits mettent moins en avant leurs caractéristiques intrinsèques que leur origine ou leur contexte d'utilisation : ils désignent dans leur définition les données personnelles des internautes ou des consommateurs et/ou les données produites, détenues, exploitées, par les entreprises. Il peut aussi s'agir de données nécessitant une expertise particulière, ou de données plus « importantes » que « grandes » (deux compréhensions possibles de l'anglais « big »). Plusieurs définitions considèrent les *big data* comme des données web, sans mentionner les données qui peuvent exister localement dans les organisations ou chez les particuliers. À la marge, quelques extraits désignent également les données textuelles, les bases de séquences ADN ou encore les données spécifiquement détenues par les GAFA (Google, Apple, Facebook, Amazon).

– **Les *big data* désignent les outils et méthodes d'analyse de données**. Une deuxième grande famille de définitions désigne non pas tant les données elles-mêmes que toutes les opérations au sens large qui leur sont appliquées. Selon ces définitions, les *big data* sont les algorithmes, les technologies, les méthodes, les applications, développés pour exploiter des données ; c'est aujourd'hui la nouvelle vague médiatique relative à l'intelligence artificielle. On trouve à ce niveau aussi bien des solutions « bas niveau » permettant des opérations basiques comme le stockage, la lecture et l'écriture de très grandes quantités de données, que des systèmes sophistiqués mobilisant des algorithmes à l'état de l'art pour des usages bien particuliers. L'espace de stockage et la puissance de calcul font partie de ces solutions au « problème » des *big data*, et représentent d'ailleurs les conditions *sine qua non* de traitements poussés des données. Cependant, pour certains, il n'est pas nécessaire que les données soient pléthoriques pour que leur utilisation entre sous l'ombrelle des *big data* ; c'est au contraire la nouveauté des techniques de manipulation de données qui va caractériser un phénomène

sociotechnique particulier. Il y a ainsi un certain nombre d'algorithmes ou de logiciels qui se voient accoler l'épithète de « *big data* » sans d'ailleurs qu'elle soit justifiée d'une façon ou d'une autre. Dans cet amas hétéroclite de techniques, deux groupes reviennent avec une force particulière : celui des outils de collecte (c'est-à-dire à la fois la collecte elle-même, mais aussi la démarche par laquelle on crée les conditions pour que la collecte soit possible) et celui des outils d'analyse, qu'ils soient artisanaux, créés *ad hoc* pour un usage très spécifique, ou industriels, fonctionnant jour et nuit en continu.

– **Les *big data* désignent les usages et applications de l'analyse de données.** La question de l'analyse conduit à celle des usages : si le développement des outils d'analyse occupe en soi les pensées et les journées de bien des ingénieurs en informatique, l'analyse se fait toujours en vue de quelque chose, une fonction ou un but qui justifie ces efforts. La recherche de régularités statistiques, de corrélations, d'ordres et de regroupements tend à mobiliser des ressources plus importantes lorsque les données sont volumineuses : aussi la destination des outils est presque toujours explicitée, car on ne peut guère se permettre de travailler en vain. En cela les entreprises, lorsqu'elles disposent de moyens importants, apparaissent comme les actrices majeures du phénomène *big data*. Elles collectent, analysent et valorisent les données de leurs utilisateurs jusqu'au point où que ces opérations constituent, pour certaines, le cœur de leur activité et leur source de création de valeur. Cette prééminence dans les *big data* du monde industriel sur le monde universitaire ne se mesure pas seulement à la quantité de données traitées, mais à leur contribution intellectuelle et technique dans la constitution de ces nouveaux outils de collecte et d'analyse et au rôle politique et social que peuvent avoir les données par son intermédiaire. Au travers des entreprises, les *big data* manifestent leur filiation avec les démarches de *business intelligence* et les outils de *business analytics*. Pour les entreprises dont l'activité ne repose pas forcément sur l'exploitation des données, cette exploitation a pour fonction de faciliter son pilotage et d'orienter les prises de décisions de ses dirigeants. Dans ces usages, les *big data* apportent à la fois de nouvelles techniques et de nouvelles manières de considérer les données et notamment leur circulation. Elles remobilisent également des notions de *storytelling* et de visualisation de données, qui inscrivent dans la problématique des données la question de leur transmission, de leur présentation. Parce que les données ne sont pas dormantes mais mobilisées au service de la gouvernance de l'entreprise, on porte en effet une attention plus grande aux formes dans lesquelles elles sont communiquées.

– **Les *big data* désignent le rôle et l'impact culturel de ces analyses**. Les *big data* caractérisent enfin une sorte d'attention portée aux données ; l'idée qu'elles sont importantes, qu'elles comptent. Mises en scène, médiatisées, décriées, elles débordent d'un cadre purement technique et fonctionnaliste pour accéder à une dimension sociale et politique. Certains extraits les caractérisent ainsi d'abord par la place qu'elles ont prises dans les sociétés, les

États et le quotidien des individus ; le fait qu'il puisse s'agir de données personnelles prend alors, notamment auprès des médias, des accents éthiques par lesquels les *big data* deviennent quelque chose dont on s'inquiète. En bien comme en mal, une mythologie se construit autour des données, qui est celle de la surveillance des individus mais aussi de la transparence des entreprises et des gouvernements. Elles désignent alors une mutation culturelle globale, au cœur de laquelle se trouvent les données et les algorithmes, et qui transforment ainsi, de façon souvent superlative, tous les aspects de la société. D'un point de vue épistémologique, on décrit un changement de paradigme qui, au-delà des outils et techniques qui émergent, définit un nouveau cadre pour l'activité scientifique et le fonctionnement des entreprises. Dans le passage du technique au culturel, les contours déjà flous de la définition se noient davantage, et les *big data* finissent par n'être plus décrites que comme un *buzzword*, procédé purement rhétorique et formule vide de sens pour laquelle on ne peut formaliser aucune définition substantielle.

Ces différentes définitions se retrouvent d'une famille d'acteurs à l'autre. Du fait que le corpus a été constitué sur un mode qualitatif, nous ne prétendons pas qu'il présente une quelconque représentativité statistique, mais seulement quelques indices de ce que les distributions effectives pourraient être. De ce fait, nous ne décrirons les quantités observées qu'avec un vocabulaire qualitatif lui aussi. Concrètement, la répartition des familles de définitions par famille d'acteurs est la suivante dans notre corpus (**Tableau 1**) :

	Données	**Outils**	**Applications**	**Rôle culturel**	**Total**
Industriel	20	10	7	6	**43**
Médiatique	16	4	1	1	**22**
Scientifique	20	17	1	4	**42**
Total	**56**	**31**	**9**	**11**	**107**

Tableau 1*. Distribution des familles de définition des big data*

Si l'on se garde, comme il se doit étant donnés les faibles volumes considérés, de formuler des considérations d'ordre statistique, on voit néanmoins que la question la moins sophistiquée, celles des données elles-mêmes, est la plus prégnante, en particulier pour les journalistes (72% du corpus journalistique, contre 52% de l'ensemble), tandis que la question des applications préoccupe plus spécifiquement les industriels (23% du corpus industriel contre 8% de tout le corpus). On retiendra surtout que toutes les définitions se retrouvent bien chez tous les acteurs mêmes si certains semblent avoir certaines prédilections. Les confirmer n'est cependant pas notre objet : nous voulions ici nous assurer que sans être homogène, la compréhension du phénomène *big data* est commune aux différents acteurs, et qu'elle s'articule autour de points d'achoppement partagés. Nous avons également pu confirmer que, parmi ces points d'achoppement, la question du volume de données (et de leurs autres caractéristiques) fait bien partie de cette compréhension mais qu'elle ne s'y limite pas.

Dans cette perspective, nous souhaitons maintenant déterminer le régime discursif de ces définitions, en établissant que ces différents points d'achoppement, que nous avons jusqu'à maintenant qualifiés de « familles de définitions », ne sont pas exclusifs ni contradictoires, et qu'il n'est donc pas légitime de les envisager en termes de valeur de vérité. Notre démarche est à la fois logique et empirique : d'un point de vue logique, nous verrons qu'il n'est pas avisé de vouloir associer un fait de langage à une définition formelle ; d'un point de vue empirique, nous allons expliciter le statut sociologique et historique qu'il faut donner à ces discours, de manière à en déduire le régime discursif.

3. Le problème logique de la définition

En examinant les discours relatifs aux *big data*, nous avons constaté qu'il n'y avait pas de définition claire et générale qui se dégageait. Les *big data* n'ont pas de traits nécessaires et suffisants qui permettraient de dessiner précisément leurs frontières. La notion ne décrit pas un objet des sciences comme l'atome décrit, si ce n'est une réalité physique, du moins un objet théorique des sciences de la nature. Ce n'est pas un concept qui subsumerait, éventuellement de manière un peu flottante, des éléments du réel, mais un phénomène culturel constitué de faits et de valeurs, de discours et de pratiques, bref d'un ensemble de croyances et de réalités qui se peuvent se prêter au regard de nombreuses disciplines comme la sociologie, l'économie, l'histoire, l'anthropologie ou encore les sciences de l'information et de la communication. Il s'agit davantage d'un mot de ralliement entre chercheurs que d'un champ de recherche en tant que tel, sauf en tant qu'objet de recherche critique : en termes d'ancrage disciplinaire, on peut travailler *sur* les *big data* mais pas *en big data*. Il ne recouvre que partiellement le champ d'action de disciplines comme la statistique ou de l'informatique) et ne se prête qu'imparfaitement à un cadre scientifique qu'il déborde largement par ailleurs.

Néanmoins, ce n'est pas parce qu'une expression n'a pas un contenu raisonné et cohérent qu'elle est dépourvue d'une existence culturelle et d'un ensemble de significations qui bien que confuses, nous permettent de nous comprendre dans l'espace du langage. De plus, force est de constater que l'absence de définition formelle caractérise la plupart des mots que nous utilisons, au moins dans le langage courant. Si j'étais, au milieu d'une phrase, sommée d'énoncer précisément le sens conféré à un mot que je venais de prononcer, je serais probablement bien embarrassée. Augustin rencontre justement cet embarras lorsqu'il cherche à penser ce qu'est le temps : « Si personne ne m'interroge, je le sais ; si je veux répondre à cette demande, je l'ignore. » Dire que nous employons des termes sans songer précisément à leur signification est aujourd'hui une banalité. Que nous ayons des difficultés à définir un terme ne veut pas donc dire que nous n'en connaissons pas le sens, mais que ce sens émerge d'un contexte, d'exemples, de situations dans lesquelles nous le savons pertinent : spontanément, nous ne savons pas toujours proposer une définition analytique, mais nous saurons toujours proposer un exemple d'utilisation des mots que nous connaissons. Même en contexte scientifique, il n'y a pas toujours de consensus entre les individus, et des notions comme celles d'énergie ou d'information

peuvent être mobilisées à travers un langage commun alors que les acteurs individuels en ont des définitions différentes, mais comparables, commensurables. Que nous ne sachions donc pas *a priori* définir analytiquement les *big data* n'en fait pas une expression vide de sens, mais bien un fait de langage soumis à l'étrange incertitude qui entoure son emploi.

En sciences cognitives et notamment d'après la théorie du prototype développée par Eleanor Rosch (1973), le contenu sémantique d'un terme n'a pas non plus besoin d'être figé autour de traits définitoires nécessaires et suffisants pour que le terme puisse être compris. Ainsi, beaucoup d'oiseaux volent mais pas tous (le manchot ne vole pas mais il est décrit vernaculairement comme un oiseau) ; en revanche tous les oiseaux ont une partie des traits considérés comme caractéristiques des oiseaux (des ailes, un bec, le fait de construire un nid, le fait de pondre des œufs, etc.). L'appartenance à la catégorie « oiseau » est graduelle, et il y a donc de « meilleurs oiseaux » que d'autres. De la même manière, la désignation « *big data* » peut être vue comme graduelle, avec des exemples plus ou moins emblématiques de cette catégorie.

La capacité à élucider le sens d'un terme a été au cœur des réflexions de la philosophie du langage du XXe siècle. Chez le premier Wittgenstein, c'est précisément le rôle de la philosophie que de procéder à cette élucidation. Dans le *Tractatus*, il écrit :

> Le but de la philosophie est la clarification logique des pensées. La philosophie n'est pas une théorie mais une activité. Une œuvre philosophique se compose essentiellement d'éclaircissements. Le résultat de la philosophie n'est pas de produire des « propositions philosophiques » mais de rendre claires les propositions. (Wittgenstein 2004)

Nous considérons ainsi que l'effort d'élucidation d'un terme est une tâche éminemment philosophique, et que l'explication des concepts peut même être considérée comme la tâche philosophique par excellence. Chez Schlick, ce travail d'élucidation est producteur de connaissances. Selon lui, toute connaissance véritable est expression, c'est-à-dire recombinaison de signes préexistants :

> Tout progrès dans la connaissance se caractérise toujours de la même manière : il s'agit de décrire une chose dans les termes de quelque chose d'autre, c'est-à-dire d'en donner une description qui consiste en une nouvelle combinaison de signes anciens. (Schlick 2003)

Un terme nouveau ne contient pas de connaissance en tant que tel : c'est sa reformulation dans des termes connus, la commensuration entre son contenu et des propositions connues, qui lui confèrent une valeur épistémique. Nous entendons mener cet effort philosophique au sujet de l'expression *big data*, afin d'élucider le sens des discours qui l'emploient, de « rendre claires les propositions » relatives aux *big data*. De fait, le *sens* des *big data*, ce qu'elles représentent en tant que phénomène culturel, est à chercher non pas tant du côté du contenu des discours, comme nous l'avons fait précédemment, mais du côté de leur statut et de leur contexte socioculturel.

4. Les *big data* comme mythologie

Si la compréhension des *big data* exprimée dans les discours que nous avons analysés ne forme pas un tout cohérent ou complémentaire de point d'achoppements, on peut néanmoins repérer une origine commune, au-delà de ces discours, à ce qui les motive. D'un point de vue historique, la cause première de tous les éléments décrits est le passage historique de l'informatique comme domaine spécialisé au numérique comme technologie généralisée. Cette généralisation recouvre à la fois une adoption par les individus, représentée par le développement de l'informatique personnelle, une transformation des institutions en tout genre et notamment des entreprises, et l'apparition d'Internet en tant que phénomène mondial massivement utilisé : nous proposons de considérer ces trois points comme les trois éléments majeurs de ce qui est communément appelé « révolution numérique ». L'avènement du numérique présente plusieurs aspects qui conduisent au phénomène *big data*, dont il est le prolongement : l'usage de ces technologies d'une part, et l'informatisation des instruments de mesure, notamment scientifiques, produisent des données sous format numérique. Le numérique rend possible le développement de nouvelles techniques de traitement de données, encouragé par leur multiplication. L'usage de ces techniques produit des résultats qui suscitent de l'intérêt et conduisent à la mise au point d'autres techniques ; ces différents points s'encouragent réciproquement et font des pratiques associées un phénomène majeur. De ce fait le phénomène attire l'attention d'acteurs plus critiques qui interrogent ses aspects juridiques, politiques, éthiques et épistémologiques, lesquels influent, par rétroaction, sur leur objet.

Le régime discursif de ce dernier type d'acteurs mérite d'être précisé. À première vue en effet, une analyse critique du phénomène *big data* pourrait apporter un éclairage pertinent pour la présente investigation ; néanmoins ce n'est pas ainsi que nous les envisageons. Dire des *big data* qu'elles sont un effet de mode, ou une construction sociale, comme le soutiennent cette typologie d'acteurs, permet indubitablement de les remettre en perspective et d'installer une distance critique avec cet objet. Cependant, le problème est que la déconstruction du phénomène *big data* n'annule pas l'existence de ses différentes incarnations matérielles, des pratiques concrètes qui y sont liées. Que leur définition soit, comme on l'a vu, problématique et que leurs frontières soient floues, ne les invalide pas en tant qu'objet, et bien au contraire dirions-nous, dans la mesure où ces incertitudes laissent précisément le jeu nécessaire à une investigation de nature philosophique. Dans *The social construction of what?* (1999), Ian Hacking examine l'insistance que l'on peut rencontrer à faire d'un objet un fait culturel relatif et contingent, comme c'est le cas dans la constructivisme sociologique. Cette insistance, généralement axiologiquement chargée, trahit un agenda normatif qui ne se contente pas d'avoir une approche critique de l'objet, mais souhaite le voir évoluer. Il écrit notamment :

> Social construction work is critical of the status quo. Social constructionists about X tend to hold that:
>
> (1) X need not have existed, or need not be at all as it is. X, or X as it is at present, is not determined by the nature of things; it is not inevitable.

> Very often they go further, and urge that:
>
> (2) X is quite bad as it is. (3) We would be much better off if X were done away with, or at least radically transformed.[2]

Du fait de leur agenda normatif, les acteurs qui s'efforcent de faire des *big data* un effet de mode artificiel font donc *de facto* partie intégrante de la controverse sur l'objet dont ils souhaitent montrer le caractère relatif. Les entreprises de déconstruction visant à instaurer une distance critique par rapport à l'objet ne se contentent pas de graviter autour de lui, mais se trouvent intégrées par lui précisément en raison de cet effort. Ces travaux, d'origine académique comme journalistique, ne sont donc pas tant à envisager comme des points de vue distanciés sur l'objet que des éléments, des parties de celui-ci. Dans le cas d'espèce, l'argument de Ian Hacking nous permet de justifier pourquoi nous les considérerons comme tels dans notre investigation.

Par l'examen des pratiques effectives auquel nous allons procéder dans les chapitres suivants, on renonce ainsi au moins temporairement à une philosophie des sciences qui se voudrait normative dans son analyse des concepts scientifiques, tout en gardant son rôle d'élucidation de ces mêmes concepts. Dans cette perspective, nous refusons l'option de proposer une définition normative des *big data* qui se surajouterait à sa mythologie et poursuivons le projet, plus ardu mais plus adéquat, de restituer pour le moment dans ses différentes dimensions la complexité du phénomène à clarifier.

Comme on l'a vu, ce phénomène est à la fois constitué de pratiques effectives, de discours relatifs à ses pratiques, et des relations systémiques entre les deux. Les acteurs des *big data* adhèrent à tout ou partie des croyances qui leur sont relatives. Dans une économie de la promesse, des discours sont formulés quant à ce qui pourrait être fait et ce qui serait rendu possible, avant que la possibilité n'ait été vérifiée empiriquement. Ces discours ne sont pas purement performatifs ni purement putatifs : ils convainquent certains acteurs, d'autres s'en emparent par tactique ; ils imprègnent les programmes de recherche et orientent les agendas universitaires. Des financements sont dégagés pour développer le sujet, mobilisant des chercheurs dont la production scientifique relève alors *de facto* du sujet. Dans l'industrie, des entreprises s'en revendiquent, modulent la notion pour qu'elle s'applique au mieux à leur activité, produisant de nouveaux discours, relayés par les médias, puis par le public. Les discours médiatiques influent à leur tour les institutions, publiques et privées, qui annoncent des politiques et des priorités en faveur du sujet, et ce sur la base de la croyance en son importance. D'autres discours adoptent un point de vue critique par rapport à cette économie de la promesse, mais pas de façon absolue, si bien que certaines hypothèses et assertions défendues dans les discours partisans sont

[2] « Le travail sur la construction social est critique du statut quo. Les constructivistes sociaux au sujet de X tendent à considérer que :
(1) X aurait pu ne pas exister, ou n'est pas nécessaire en tant que tel. X, ou X tel qu'il se présente aujourd'hui, n'est pas régi par la nature des choses; il n'est pas inévitable.
Très souvent, ils vont plus loin et soutiennent que :
(2) X est très mauvais en tant que tel. (3) Il vaudrait mieux que se débarrasser de X, ou du moins le transformer radicalement. » (notre traduction)

maintenues dans ces discours critiques. Parmi ces critiques on rencontre à la fois des chercheurs extérieurs au sujet, qui en font un objet de recherche et deviennent de ce fait partie intégrante de cette économie, et des praticiens des *big data* quels qu'ils soient, qui sont à même de formuler un certain nombre de critiques plutôt internalistes quant aux limites de leurs propres pratiques. Tous ces discours forment un système de croyances qui n'est pas une simple boucle de circulation d'information, mais un réseau complexe d'échanges d'idées qui existent sous un régime discursif rhétorique et axiologique plutôt qu'argumentatif et véridictionnel.

Sous cet angle, le phénomène *big data* relève davantage de l'anthropologie culturelle que de la philosophie des sciences, bien que ses *leitmotivs* rhétoriques se matérialisent aussi sur des canaux comme les revues scientifiques qui sont les voies habituelles d'expression des discours scientifiques.

Nous proposons à ce titre de mobiliser le terme barthésien de *mythologie*, c'est-à-dire un discours non plus descriptif mais idéologique, axiologiquement chargé, alimenté par des représentations collectives « maintenues stagnantes dans l'erreur par le pouvoir, la grande presse et les valeurs d'ordre » (Barthes). Quoique la définition de Barthes soit elle-même axiologiquement chargée, puisqu'elle vise à dénoncer l'emprise de la bourgeoisie de son époque, cette ambition dénonciatrice lève toute ambiguïté quant au caractère doxastique des discours mobilisant la notion de *big data*, ou autrement dit, met en suspens un éventuel caractère véridictionnel.

Afin de mettre en évidence cette nature doxastique, nous analysons quelques procédés rhétoriques utilisés : pour cela nous nous concentrons sur quelques ouvrages emblématiques de la mythologie des *big data*. Nous extrayons de ces ouvrages quelques thèses d'ordre épistémologique relatives aux *big data* et les mettons en regard d'autres discours sur les pratiques techniques et scientifiques, ceux de la tradition épistémologique. Nous ne mobiliserons pas ces discours pour les prendre à notre compte (c'est-à-dire indépendamment du fait qu'on puisse effectivement les considérer comme valables dans notre contexte), mais de manière instrumentale, pour mettre en relief le caractère doxastique des thèses d'ordre épistémologique identifiées dans les *big data*.

5. Les mythes fondateurs

À travers notre premier corpus d'une centaine de définitions, nous avons opté pour une approche prudente, qui ne posait aucune signification préalable au terme de « *big data* » et s'efforçait d'en écouter tous les discours. Nous n'avons pas non plus étudié le ton, le style, les tropes rhétoriques de ces définitions. Le résultat de cette analyse est qu'il existe autant de significations que de discours, mais que l'on peut leur néanmoins en tirer des traits communs, un contenu sémantique relativement partagé. Le phénomène des *big data* prend notablement la forme d'un imaginaire, qu'il est difficile de circonscrire car proposer (ou réfuter) une définition de ce que sont les *big data* nous fait devenir un acteur de cet imaginaire. Les *big data* ne sont cependant pas qu'un objet sémiotique ancré dans une

formule : il s'agit également d'un certain nombre de principes, de croyances – fondées ou non – qui reflètent une partie de la mentalité d'une époque. Ancré dans le temps, ce phénomène présente des caractéristiques qui ne sont nécessaires ni suffisantes, bien qu'il soit presque toujours question de données numériques contemporaines. De ce point de vue, d'autres discours que ceux qui emploient l'expression littérale peuvent être reliés au phénomène.

Dans cette perspective, nous avons sélectionné cinq documents emblématiques (quatre livres et un article de presse) qui nous serviront de base empirique pour examiner les thèses d'ordre épistémologique du phénomène *big data*. Ces documents relaient et reflètent particulièrement bien l'imaginaire qui leur est associé. Deux d'entre eux, les suivants, y ont même largement contribué :

- « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », article de Chris Anderson, alors rédacteur en chef du magazine *Wired*, paru en 2008 ;
- *Big data: A Revolution That Will Transform How We Live, Work, and Think*, livre de Viktor Mayer-Schönberger, professeur de droit à Oxford, et Kenneth Cukier, journaliste, paru en 2013.

Trois autres ouvrages moins connus mais néanmoins emblématiques ont retenu notre attention :

- *The Fourth Paradigm*, édité par Tony Hey, Stewart Tansley et Kristin Tolle, travaillant tous les trois à des postes de direction au département recherche de Microsoft et publié par Microsoft Research en 2009 ;
- *The Signal and the Noise. Why So Many Predictions Fail but Some Don't*, livre du statisticien et journaliste Nate Silver, paru en 2012 ;
- *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, publié en 2015 par Pedro Domingos, professeur d'informatique à l'Université de Washington.

Dans l'ensemble, ces ouvrages ont une notoriété significative dans le monde universitaire, à l'exception du dernier, sans doute trop récent à l'heure où nous écrivons ces lignes (**Tableau 2**).

Aucun de ces documents n'est un ouvrage de recherche à proprement parler, bien que certains de leurs auteurs aient une activité universitaire et publient régulièrement par ailleurs dans des revues scientifiques. À nouveau, le fait d'étudier le phénomène *via* une littérature extrascientifique en dit déjà quelque chose dans la mesure où les publications scientifiques qui traitent explicitement des *big data* présentant plutôt des analyses souvent critiques du phénomène culturel (autrement dit, de la mythologie) que des rapports sur des pratiques effectives d'analyse de données. Si l'on cible maintenant, dans la littérature scientifique, toutes les publications mobilisant de l'analyse de données et présentant un certain nombre de caractéristiques de la mythologie des *big data*, on se retrouve face à un volume de publications à étudier difficilement abordable, et qui serait par ailleurs incomplet. Il

faudrait notamment leur adjoindre des rapports produits par des entreprises ou des institutions, des documents internes, des publications en ligne, des articles de presse, bref toute la diversité de contenus que l'on a aperçue en travaillant sur notre premier corpus. Ici, nous cherchons pour ainsi dire à remonter aux sources de la mythologie, à identifier les termes dans lesquels elle se pose, à la fois dans sa complexité et à travers quelques traits saillants et emblématiques.

Ouvrage	**Nombre de citations**[3]
(Anderson 2008)	1243
(Hey, Tansley et Tolle 2009)	2090
(Silver 2012)	887
(Mayer-Schönberger et Cukier 2013)	3187
(Domingos 2015)	95

***Tableau 2**. Nombre de citations académiques des ouvrages*

Dans ce but, nous avons privilégié des essais et ouvrages de vulgarisation qui se prêtent bien à l'exposition d'une position relative à un sujet. Aucun n'est un ouvrage critique par rapport au phénomène des *big data*, et nous considérerons dans l'ensemble que les auteurs soutiennent une posture technophile à ce sujet. C'est d'ailleurs, en quelque sorte, l'ingéniosité de ces textes que de propager leurs positions en se présentant comme des ouvrages de vulgarisation, censément descriptifs d'un état de fait : les points de vue soutenus sont perçus comme neutres, et relayés comme tels. Nous y avons retrouvé tous les traits repérés dans notre premier corpus, ainsi qu'une tendance à être repris et formulés avec une viralité remarquable, notamment en ce qui concerne l'article de Chris Anderson. Tout en y étant largement contesté, cet article est systématiquement cité dans tous les articles de recherche qui revendique une posture critique vis-à-vis du phénomène *big data*.

Chacun de ces ouvrages expose un certain nombre de thèses qui sont, non seulement en faveur, mais particulièrement emblématiques de l'imaginaire des *big data*. Beaucoup de ces thèses ont trait à la pratique de la science et à la façon dont l'abondance de données modifierait ces pratiques ; plus largement, elles portent sur nos modalités de connaissance du monde, et l'impact plus général qu'elles peuvent avoir. Néanmoins, il ne s'agit pas tant dans l'ensemble de thèses épistémologiques que d'une forme de foi dans les *big data*, c'est-à-dire d'une croyance qui n'a pas besoin de faire l'épreuve du réel pour être maintenue par le sujet : comme on le verra, les exemples proposés ne jouent pas leur rôle de preuve empirique dans la mesure où ils sont mal décrits, mal reliés à la thèse défendue d'un point de vue logique, et trop anecdotiques ou singuliers pour faire l'objet d'une montée en généralité. C'est aussi pour cela qu'ils ont bien le statut de mythe, c'est-à-dire d'un discours fictionnel tenu pour vrai.

[3] Chiffres Google Scholar, novembre 2017.

6. Les composantes du mythe

Sur le plan stylistique, chacun des documents pourrait donner lieu à une analyse littéraire et rhétorique approfondie, et des critiques détaillées : les formes de raisonnements, les tropismes, les références communes aux différents documents, sont autant d'expressions de l'imaginaire des *big data*, que nous ne pourrons détailler ici. Par exemple, David Beer (2016) a analysé le tropisme de la vitesse et de l'instantanéité chez les fournisseurs de solutions d'analyse de données. Il écrit ainsi :

> A number of themes emerge when exploring the envisioning of the power of data analytics, but speediness, immediacy and the promises of real-time knowing are particularly prominent. These are products and solutions accompanied by claims such as: 'The fastest easiest way to understand your data' (Appendix 1, ref 3), or 'Fast analytics for everyone' (Appendix 1, ref 15), or 'Fast-cycle Business ready insights on more data' (Appendix 1, ref 17) and 'We provide the world's fastest, easiest, and most secure data platform' (Appendix 1, ref 22). Speed dominates. There is a sense here of a profound need for acceleration. Effective and productive data analytics are seen to be fast data analytics. Good analytics are those that are seen to produce instant results.

Dans le même esprit, Turow, Mcguigan et Lee (2015) montrent comment les entreprises qui ont développé l'analyse des données de leurs clients tendent à propager l'idée que ces pratiques sont un aspect naturel de la vie, quelque chose de banal et évident, dans le but de les faire accepter par les consommateurs. Dans les ouvrages que nous avons sélectionnés, deux schèmes rhétoriques ressortent de manière particulièrement évidente :

- L'**hybris de l'apprenti-sorcier** : en termes d'acteur, tous les documents pointent vers l'informaticien ou le statisticien comme figure de la pratique des *big data*. Néanmoins, le domaine d'intervention des *big data* déborde de celui de l'informatique ou de la statistique, dans la mesure où il ne s'agit pas seulement d'un art du calcul, mais d'un mode d'accès au réel, qu'il s'agisse de le représenter ou d'y intervenir, et qu'il soit physique ou surtout social. Face à cela, l'informaticien s'appropriant un domaine nouveau est pris d'une forme d'hybris où il se prend littéralement pour Dieu :

  > A programmer—someone who creates algorithms and codes them up—is a minor god, creating universes at will. You could even say that the God of Genesis himself is a programmer: language, not manipulation, is his tool of creation. Words become worlds. Today, sitting on the couch with your laptop, you too can be a god. Imagine a universe and make it real. The laws of physics are optional. (Domingos 2015)

- Le **vertige de l'énumération** : le volume de données est fréquemment présenté comme une caractéristique essentielle des *big data*. Néanmoins il convient de constater qu'il n'apparaît pas seulement comme une vertu épistémique, telle que la fiabilité ou la représentativité qui en découlerait. Ce volume est surtout décrit sous l'angle de la quantité physique qu'il représente (l'espace de stockage informatique qu'il occupe) et non de la

richesse de son contenu sémantique : il y a un glissement de sens entre information et connaissance. Dans notre corpus, il occupe ainsi une fonction rhétorique, celle d'impressionner le lecteur et de lui donner l'impression qu'il est face à quelque chose de l'ordre du sublime, qui dépasse l'entendement. Les déclinaisons des unités de mesure de l'information numérique (gigabit, pétabit, exabit, etc.), les énumérations de comparaisons, les commensurations purement quantitatives avec des masses d'information analogiques, en particulier les livres imprimés, induisent l'idée que la quantité est bonne en soi, et qu'en l'occurrence on est face à des quantités qu'on ne peut appréhender intellectuellement, mais sur le mode du vertige :

> There is no good way to think about what this size of data means. If it were all printed in books, they would cover the entire surface of the United States some 52 layers thick. If it were placed on CD-ROMs and stacked up, they would stretch to the moon in five separate piles. In the third century B.C., as Ptolemy II of Egypt strove to store a copy of every written work, the great Library of Alexandria represented the sum of all knowledge in the world. The digital deluge now sweeping the globe is the equivalent of giving every person living on Earth today 320 times as much information as is estimated to have been stored in the Library of Alexandria. (Mayer-Schönberger et Cukier 2013)

Au-delà de la forme au moins partiellement argumentative de ces ouvrages, on constate donc qu'ils soutiennent des discours axiologiquement chargés, appuyés par des croyances présentées comme ayant un caractère d'évidence, et des *topoi* rhétoriques qui visent à convaincre le lecteur par l'adhésion à un imaginaire plus que par des arguments rationnels. Mettre le doigt sur ces procédés est une nécessité préalable à l'exercice de démêlage entre les procédés rhétoriques de promotion de la mythologie des *big data*, les valeurs les contenus déontiques qu'elle véhicule, et enfin les prétentions épistémologiques de cette mythologie.

Dans les ouvrages sélectionnés, les promesses des *big data* sont formulées de manière absolue, soit comme quelque chose qui va se produire inévitablement, soit comme quelque chose qui s'est déjà produit de toute façon. Ce ne sont pas des hypothèses mais des affirmations. Les postures sont révolutionnaires et non réformistes : il ne s'agit pas d'introduire des inflexions dans les méthodes, les outils, les contenus théoriques, de faire évoluer les pratiques de manière incrémentale, mais de « tout changer ». Ainsi, Anderson soutient que les *big data* rendent « la méthode scientifique » obsolète et que les sciences, notamment les sciences sociales, vont disparaître. Mayer-Schönberger et Cukier présentent les *big data* comme une révolution qui va transformer notre façon de penser. Domingos écrit que son algorithme va changer le monde. Le ton de ces textes a quelque chose d'eschatologique ; tout se passe comme s'il fallait faire table rase du vieux monde pour laisser place au nouveau.

Comme souvent dans les discours révolutionnaires, on rencontre également une réécriture rétrospective de l'histoire par laquelle les *big data* apparaissent comme plus anciennes qu'elles ne le

sont, ce qui est censé les légitimer. Ainsi, la constitution des *datamarts* par les distributeurs tels que Walmart dans les années 1980, ou les pratiques de modélisation et simulation informatique en physique et en biologie depuis les années 1960, sont présentées rétrospectivement comme des pratiques qui relèvent des *big data*. L'histoire même des statistiques depuis le XVIIe siècle est réécrite téléologiquement comme une ascension irrésistible vers l'avènement des *big data*. Nate Silver notamment se présente comme un statisticien et décrit une continuité entre les statistiques des siècles précédents et l'analyse de données contemporaines. Des références à l'histoire antique et médiévale, notamment à la bibliothèque d'Alexandrie, inscrivent les *big data* dans un temps long et les présentent comme un retour aux sources après un temps d'obscurantisme. On assiste ainsi à un paradoxe selon lequel les *big data* ont toujours été là, sont advenues progressivement et constituent en même temps une nouveauté radicale.

L'originalité de ces discours sur la fin de la science est aussi qu'il s'agit cependant de discours scientistes, qui espèrent rendre la science meilleure (grâce aux *big data*) et non la détruire ; la fin de la science doit se produire pour que la science puisse renaître. Il se peut que les disciplines scientifiques historiquement constituées disparaissent, mais le savoir scientifique pour sa part ne s'en portera que mieux. Les accents prophétiques de ces textes ne reflètent pas une intention de renverser la science de son piédestal pour propager autre chose, par exemple une doctrine religieuse antiscience, mais bien de renforcer la foi dans le savoir scientifique en annonçant l'avènement de « la vraie science ».

L'analyse computationnelle de données massives et ses algorithmes apparaît ainsi comme une amélioration par rapport aux méthodologies scientifiques existantes, qui n'auraient plus lieu d'être. Cependant, tous les documents ne défendent pas nécessairement un remplacement de la science actuelle par les *big data* ; il semble que celui-ci va se produire dans tous les cas, mais ce n'est pas forcément un but en soi. Pour emprunter la terminologie des doctrines religieuses, on rencontre à la fois des fondamentalistes, qui veulent remplacer l'état de fait par leur propre doctrine, et des réformistes qui revendiquent la liberté de pratiquer leur propre foi à côté de la foi officielle. Dans l'ensemble, les auteurs semblent plutôt considérer la science « traditionnelle » comme un vieux monde devenu obsolète et en attente de réforme que comme des pratiques inorthodoxes qui doivent être éliminées par l'inquisition ; que le nouveau monde se construise à côté ou à la place de l'ancien ne semble pas être une différence cruciale pour les acteurs étudiés.

Par bien des aspects, la mythologie des *big data* n'est pas sans rappeler le positivisme scientifique d'Auguste Comte, qui présente lui aussi des accents religieux (au point d'avoir d'ailleurs donné naissance au positivisme religieux) et déborde de la réflexion sur la connaissance pour s'imposer comme doctrine sociale. Comme le positivisme comtien, les *big data* présentent une vision totalisante du savoir scientifique, fortement empiriste et mécaniste, et qui se présente davantage sous la forme d'un ensemble de croyances relatives à ce que la science pourrait être, qu'à une description empirique

des théories et pratiques scientifiques. L'empirisme des *big data* rejette les cadres théoriques à la façon dont Comte rejette la métaphysique ; les deux mythologies attribuent une valeur positive, inconditionnelle, à la donnée, et rejettent la recherche d'explications causales.

Ces composantes mythologiques confirment que notre hypothèse de départ doit être reformulée : les discours des *big data* ne sont pas à envisager seulement comme une technologie, une mise en raison des savoir-faire artisanaux qui viserait à les organiser et les rendre transmissibles, mais aussi comme une mythologie qui s'emploie à légitimer ces pratiques, à les faire accepter, en propageant un certain nombre de croyances positives relatives à l'analyse computationnelle de données. Ayant mis en évidence la forme rhétorique de ces croyances, nous analysons maintenant leur contenu, à partir de quelques exemples dont nous expliciterons le statut.

Chapitre II.
Un imaginaire épistémologiquement chargé

Sur le plan du contenu et des assertions, la mythologie des *big data* s'appuie sur un certain nombre d'oppositions entre l'ancien et le nouveau, et tout particulièrement entre la science traditionnelle et les *big data*. De ce point de vue, les difficultés que nous avions à envisager le phénomène au prisme de la philosophie des sciences se résorbent ici, puisqu'en ces termes nous pouvons en effet, comme le propose Schlick « décrire une chose dans les termes de quelque chose d'autre », et plus précisément décrire les *big data* telles qu'elles sont envisagées au sein du mythe à l'aune de la science traditionnelle.

Notons ici que le fait même de parler de « la science » traditionnelle comme un fait monolithique est quelque chose que nous retrouvons dans le mythe des *big data*, mais qui, en soi, est discutable et discuté en philosophie des sciences. Nous verrons d'ailleurs que certaines oppositions proposées par le mythe ne sont plus tenables précisément dès lors que l'on envisage les sciences comme un ensemble hétérogène avec des spécificités disciplinaires, méthodologiques, etc. Néanmoins, pour ne pas complexifier l'analyse dans un premier temps, nous allons exposer telles quelles un certain nombre de ces oppositions (**Tableau 3**) que nous allons développer.

	Science traditionnelle	***Big data***
Visée	Expliquer	Prédire
Fondement	Modèles	Données
Données	Échantillon représentatif	Grands volumes hétérogènes
Objets	Causes	Corrélations
Raisonnements	Déduction	Induction
Acteurs	Universités	Entreprises

Tableau 3*. Récapitulatif des principales oppositions.*

Ces caractéristiques, qui n'ont pas l'ambition d'être exhaustives, ne sont pas non plus systématiquement présentes : elles sont de l'ordre des traits d'appartenance à une catégorie au sens

de la théorie du prototype. Néanmoins elles permettent de brosser à grands traits un portrait des thèses des *big data* d'ordre épistémologique telles qu'elles sont défendues et critiquées. Dans les pages qui suivent, nous allons examiner ces caractéristiques, et les opposer à la vision plus traditionnelle des régimes de production de connaissances dans l'histoire des sciences et des idées. En contournant l'imaginaire eschatologique qui leur donne une charge rhétorique envahissante, nous les examinons comme des assertions d'ordre épistémologique que l'on peut mettre en dialogue avec les assertions classiques de l'histoire et de la philosophie des sciences. Il s'agit ainsi de procéder à une analyse critique de manière à les contextualiser, les détailler, les exemplifier, en préciser le domaine, bref les rendre claires et en expliciter le sens dans des termes plus caractéristiques de la philosophie des sciences. Une fois clarifiées, on peut en apprécier le contenu épistémologique, et le discuter par analyse logique et à travers des exemples issus de l'histoire des sciences, cités dans ces discours, ou encore se revendiquant de la sphère des *big data*.

Avant cela, soulignons que certaines des thèses soutenues se retrouvent occasionnellement dans un certain nombre de projets connus et considérés généralement comme des projets *big data*, les justifiant en apparence. Néanmoins, cette adéquation entre les discours et les pratiques s'accompagne, dans les exemples qu'on va voir, d'un certain nombre d'obstacles à l'intelligibilité des données. En d'autres termes, l'application littérale de ces thèses nuit aux pratiques d'analyse de données et à la qualité des résultats obtenus. La mobilisation de ces exemples tend plutôt à affaiblir les thèses qui les mentionnent comme des preuves. Ces exemples de projets sont néanmoins éclairants, à la fois en tant que tels, pour illustrer notre objet, et par la façon dont ils sont saisis par les mythes épistémologiques des *big data*.

Un projet en particulier ressort des différents discours : cité notamment par Domingos, et par Cukier et Mayer-Schönberger, il s'agit du projet Google Flu Trends, qui agrège l'ensemble des traits considérés comme propres aux *big data*. Il peut être vu (et est d'ailleurs souvent présenté) comme un projet emblématique de ce type d'approche. L'ambition du projet Google Flu Trends était de prédire les épidémies de grippe mieux que la principale agence gouvernementale de santé aux États-Unis (les *Centers for Disease Control and Prevention*, dits CDC), s'appuyant pour cela sur l'analyse des historiques de recherches de symptômes sur son moteur de recherche (Ginsberg et al. 2009). L'un des enjeux du projet étant d'anticiper davantage les épidémies, les chercheurs de Google n'utilisaient pas les données des CDC pour faire leurs prédictions, car elles sont produites et diffusées moins instantanément que les historiques de recherche, auxquels Google a un accès immédiat. Ces « *big data* » n'étaient donc pas combinées à des données plus classiques, mais comparées avec celles-ci pour valider les prédictions. Un problème épistémologique plus subtil est que les données des CDC sont utilisées comme étalon, c'est-à-dire que, pour s'assurer que les prédictions du projet Flu Trends sont justes, elles sont comparées aux données des CDC. Cependant ces données elles-mêmes ne sont

qu'une mesure du fait empirique (l'épidémie), et non le fait lui-même : que les prédictions de Google et ces données soient concordantes ne veut pas nécessairement dire qu'elles sont exactes.

D'un point de vue institutionnel, c'est un projet mené par une entreprise privée du secteur informatique, en l'occurrence l'un des GAFA, qui entend faire mieux que les dispositifs déjà en place (sur le plan de la rapidité à obtenir un résultat plutôt que sur la précision du résultat) : la différence d'acteurs entre *big data* et science traditionnelle est confirmée dans ce cas. Il s'appuie sur des masses de données qui n'avaient jamais été utilisées de cette façon dans la perspective d'une analyse de corrélation. Plutôt qu'une recherche de compréhension de la propagation des épidémies de grippe à travers une modélisation épidémiologique, il considère que les corrélations sont suffisantes ; il ne viserait donc pas à expliquer mais à prédire. Nouvelles données, approche corrélative, inductive et prédictive, sont bien les traits que nous avons dégagés du mythe des *big data*.

Le projet Google Flu Trends est aujourd'hui inactif ; il s'est avéré que ses prédictions en matière de volume de consultations médicales annonçaient occasionnellement le double de ce qui était enregistré par les CDC et fournissait ainsi de moins bons résultats que les analyses faites à partir des données des CDC (Lazer et al. 2014). Dans l'analyse de l'échec du projet, c'est explicitement son approche *big data* qui est invoquée comme explication, et notamment le fait de ne pas avoir exploité les données « classiques » des CDC. Lazer et ses collègues dénoncent ainsi l'hybris des *big data*, défini comme

> the often implicit assumption that *big data* are a substitute for, rather than a supplement to, traditional data collection and analysis.

La prétendue supériorité des *big data* par rapport aux « small data » est ainsi remise en cause par leur imprécision, les biais qu'elles introduisent, et leur caractère instable du fait que leurs paramètres de constitutions sont instables et non maîtrisés (les algorithmes de Google changent, ce qui introduit une discontinuité dans les données produites par ces algorithmes).

De manière plus générale, il n'est pas rare en effet que les exemples emblématiques des *big data* se révèlent être des échecs du point de vue des résultats produits, de la véracité des prédictions qu'ils formulent. Par exemple, le statisticien Nate Silver, qui avait prédit avec succès le résultat des élections américaines en 2012, donnait Hillary Clinton gagnante en 2016[4]. Le caractère prédictif d'autres sources de données dites « *big data* », comme par exemple celui de Twitter concernant les élections, a été largement remis en cause (Gayo-Avello 2012). Dans l'ensemble, l'utilisateur moyen des technologies de recommandation algorithmique, à l'œuvre chez Amazon, Netflix ou Facebook, reste sans doute plus marqué par les difficultés que rencontrent ces systèmes à faire des propositions pertinentes.

[4] http://projects.fivethirtyeight.com/2016-election-forecast/ (consulté le 16 novembre 2017)

Qu'il y ait occasionnellement ou même souvent, dans toute forme d'entreprise intellectuelle, des résultats décevants, pourrait être plutôt le signe de leur honnêteté que de leur invalidité : pour le dire en des termes poppériens, qu'elles se trompent parfois montre qu'elles sont falsifiables et non dogmatiques. Il y a cependant, d'un point de vue argumentatif, un affaiblissement du raisonnement du fait que sont érigés en exemple emblématique des projets jugés moins réussis que ce qu'ils ont vocation à remplacer, et sont présentées comme triomphantes des techniques qui doivent encore faire leurs preuves. Du point de vue de la rhétorique qui mobilise ces exemples, ils ne jouent pas leur rôle de preuve, et invalident plutôt les arguments qui s'appuient sur eux. Dans notre perspective de production de connaissance à partir des *big data*, on retient ainsi que l'approche défendue n'est pas soutenable, ou du moins insuffisante, pour parvenir à des résultats satisfaisants.

À l'inverse, d'autres anecdotes, comme celle de l'entreprise de distribution qui aurait prédit la grossesse d'une cliente avant que son père n'en soit informé, ont eu un retentissement médiatique important, et ont été des éléments de preuve plus convaincants de la capacité prédictive de systèmes automatisés. Néanmoins, ces anecdotes restent relativement rares. Cette rareté a une certaine signification pour un système prédictif. En effet, un système prédictif qui doit choisir entre plusieurs propositions aura des probabilités très élevées de formuler une bonne prédiction une partie du temps : comme la montre arrêtée de Gettier affiche la bonne heure deux fois par jour, un système prédictif même complètement aléatoire fera une bonne prédiction de temps en temps. La médiatisation de ces bonnes prédictions (catalysée de plus par les enjeux éthiques que soulèvent la révélation d'une grossesse pas encore annoncée) accentue un biais de confirmation quant à l'efficacité des systèmes prédictifs.

On notera par ailleurs que de tels accidents, de l'ordre de la bonne réponse involontaire, ne sont pas inhabituels en informatique au sens large : il arrive par exemple qu'un bon résultat soit dû à deux erreurs successives qui s'annulent, et non à la validité d'une démarche d'analyse intentionnelle. Pour prendre un exemple historique un peu plus sophistiqué, la célèbre victoire aux échecs de l'ordinateur Deep Blue face à Kasparov serait en fait due à une erreur de programmation, qui conduisit Deep Blue à jouer un coup aléatoire, ce qui déstabilisa Kasparov et lui fit perdre la partie : ici ce n'est pas le choix fait par la machine, mais la croyance de Kasparov dans la justesse de ce choix, qui lui a assuré la victoire.[5] De manière générale, les exemples positifs comme négatifs retenus par les médias (et par les documents que nous avons sélectionnés) sont les plus frappants, ceux qui ont la plus grande force rhétorique : les erreurs marquantes, les succès surprenants. Néanmoins cela ne nous renseigne pas sur l'efficacité générale des techniques employées (par exemple, le taux de succès d'un système prédictif) : d'un point de vue argumentatif, la référence à de tels exemples ne permet ni d'infirmer ni de confirmer l'extraordinaire succès attribué aux *big data*.

[5] L'origine de cette précision est l'un des ouvrages de notre corpus, le livre de Nate Silver, où il rapporte une conversation avec Murray Campbell, qui dirigeait l'équipe en charge de Deep Blue dans les années 1990.

Nous avons donc vu que certaines caractéristiques attribuées aux projets *big data*, comme le caractère prédictif, la nouveauté des données, etc., sont bien présentes dans ces projets, mais que ces projets en eux-mêmes ne peuvent jouer leur rôle de preuve empirique et ne démontrent pas le progrès que les *big data* apporteraient par rapport aux sciences traditionnelles. En d'autres termes, l'affirmation selon laquelle on peut par exemple se passer de modèle théorique est affaiblie par le fait que, lorsque c'est le cas, les résultats ne sont pas satisfaisants[6] y compris dans les exemples avancés par les défenseurs des *big data*.

Nous allons maintenant examiner successivement chacune de ces affirmations. Nous verrons ainsi que ces oppositions entre *big data* et sciences traditionnelles sont en réalité, pour la plupart, des discussions qui ont lieu au sein de la philosophie des sciences. Elles polarisent dans des termes extrêmes une réalité plus complexe, qu'il nous appartient de restituer ici, en termes de pluralité de situations possibles et de points de vue sur ces situations, de manière à en retirer des traits pertinents pour une démarche de production de connaissances à partir de données massives.

1. La thèse du renouvellement des acteurs de la production de connaissances

Dans l'imaginaire des *big data*, les exemples proposés sont presque toujours des exemples industriels. Ils proviennent généralement des grandes entreprises informatiques (Microsoft, IBM, etc.) ou de la Silicon Valley : Google, qui est les deux à la fois, est le plus emblématique et le plus fréquent. Comme le souligne Sabina Leonelli (2014) :

> it is no coincidence that most of the examples given by Mayer- Schönberger and Cukier come from the industrial world, and particularly globalized retail strategies as is the case of Amazon.com

Dans l'exemple de Google Flu Trends, l'entreprise et les données de masse ont remplacé les institutions traditionnelles. De la même manière que les *big data* auraient remplacé la science, les entreprises du numérique se seraient substituées aux universités, aux laboratoires et autres institutions qui produisaient les connaissances scientifiques du « vieux monde » telles que le *Centers for Disease Control and Prevention*. Si d'une part il y a en effet une diversification des acteurs épistémiques dans les *big data*, en revanche les institutions scientifiques traditionnelles n'ont nullement disparu. Le livre de Hey, Tansley et Tolle fournit à ce titre une liste foisonnante des disciplines concernées par ce qu'ils appellent la *data-intensive science*. On y retrouve notamment l'astronomie, la génomique, la climatologie, l'océanographie ou encore la neurobiologie. Certains chercheurs travaillant sur les *big data* sous l'angle de l'épistémologie ou des STS, comme justement

[6] Nous discutons ci-dessous le fait même que les *big data* se passent de modèle théorique.

Sabina Leonelli, ont concentré leurs travaux sur l'épistémologie des pratiques effectives des *big data* à ces disciplines scientifiques.

La mutation de ces disciplines est indubitable : elle se caractérise par l'utilisation croissante d'instruments de mesure informatisés : télescopes, microscopes, capteurs, ne sont plus des instruments qui permettent de « voir à travers », d'observer indirectement, mais des ordinateurs qui observent le monde, puis produisent et traitent des données numériques.[7] Cependant, une science ne s'évalue pas seulement à l'aune de ses instruments de mesure. Il est factuellement vrai que la pratique scientifique consiste, au moins en partie, à « collecter, sélectionner et analyser des données », comme pourrait le confirmer un sociologue des sciences radicalement externaliste, qui se contenterait d'observer l'activité d'un laboratoire. Néanmoins cette vision purement empirique des pratiques est aussi incomplète que de dire que le chant lyrique consiste à bouger les lèvres et expulser de l'air. Elle passe à côté des principes, des méthodes, des enjeux, bref de tous les objets classiques de la philosophie des sciences. De ce fait, si le livre annonce un « nouveau paradigme », empruntant ici la terminologie kuhnienne, il n'est pas étonnant que l'exposition des nouvelles normes apportées par ce paradigme soit moins limpide : les *data-intensive sciences* se présentent comme une voie d'accès à l'expérience qui ne porte donc qu'anecdotiquement sur les théories qu'elle pourrait éventuellement modifier, et donc sur le paradigme qui régit ces théories. Elles relèvent de la méthode et de l'expérimentation plus que de la théorie et du paradigme scientifique Il y est question tout à la fois de nouvelles méthodes d'analyse de données, de reproduction de la recherche, de problèmes de stockage des données et de puissance de calcul, de collaboration avec la recherche en informatique, etc. Les *data-intensive sciences* apparaissent tout aussi bien constitutives de la pratique scientifique (ce *sont* les sciences) qu'au service de celle-ci (c'est un *outil* au service des sciences). On retrouve ici la variabilité des usages d'une même notion (et ce au sein d'un seul ouvrage, fût-il collectif), et dont on retiendra provisoirement qu'en matière de lieux de pratique des *big data*, plusieurs disciplines scientifiques sont effectivement concernées, sous diverses modalités.

Néanmoins, le phénomène *big data* est aussi, et peut-être surtout, un phénomène extrascientifique, non pas au sens où l'analyse computationnelle de données massives serait un phénomène essentiellement industriel, mais en ceci qu'il ne se fixe pas sur un lieu précis. C'est même sans doute un trait caractéristique de son imaginaire que de ne pas thématiser la différence entre université et industrie. Pour le contexte de la recherche en sciences sociales, Noortje Marres (2012) suggère qu'il y a, avec le numérique, une redistribution des acteurs qui s'accompagne d'un déplacement de la capacité

[7] A noter que ce style de pratique scientifique n'est pas ce que décrit Jim Gray, décédé avant la publication de l'ouvrage qui lui est dédié. Bien qu'il ne fasse pas partie des contributeurs du *Fourth Paradigm* à proprement parler, la transcription d'une de ses conférences fait figure d'introduction ; il apparaît tout au long de l'ouvrage comme une figure de référence. Pour lui, cette science computationnelle qui s'appuie de façon majeure sur l'analyse de données n'est encore que le troisième paradigme (après la science empirique et la science théorique) ; le quatrième serait la eScience qui serait la réconciliation des trois. Néanmoins d'un chapitre à l'autre le quatrième paradigme n'est pas le même, ce pourquoi nous ne pouvons pas détailler les différents sens sans entrer dans une analyse spécifique de l'ouvrage, qui n'est pas notre objet.

de recherche hors du monde universitaire, et d'une renégociation de la division du travail : le monde universitaire ne disparaît pas du champ, mais voit son rôle réécrit à travers les innovations qui le traversent, sans que ce soit d'ailleurs une nouveauté liée au phénomène spécifique des *big data* :

> For a long time already, academics have not been the only or even the main protagonists of research, as other actors have historically played active roles in the production of knowledge (Latour, 1988; Law, 2004). It is just that the conventional understanding of science and innovation makes it difficult to acknowledge the active contributions of 'non-scientists' as meaningful contributions to research and innovation, without problematizing the status of our knowledge. (Marres 2012)

La prééminence des acteurs industriels dans la mythologie des *big data* n'est pas une revendication politique à supprimer les barrières entre l'université et l'industrie, ni une comparaison entre les deux mondes visant à montrer la supériorité de l'un ou de l'autre, mais une mise en évidence de la redistribution. Les résultats industriels ne sont pas présentés comme particulièrement plus scientifiques que ceux des universités, et la question de la performance et de l'efficacité des entreprises n'est pas thématisée de cette façon-là ; de fait, les savoirs produits peuvent être le résultat d'une collaboration entre les mondes universitaires et industriels, auquel cas l'opposition institutionnelle n'a pas de sens. Il s'agit davantage d'une absence de problématisation. L'opposition se joue conceptuellement entre la science traditionnelle et les *big data*, mais pas institutionnellement entre l'université et l'industrie : d'un point de vue socio-économique, il y a une variabilité dans la façon dont les acteurs contribuent au phénomène *big data*, mais cette variabilité n'occasionne pas une disparition des acteurs traditionnels.

2. La thèse de la fin du régime explicatif

Dans *The Fourth Paradigm*, il n'est pas question en effet pour les acteurs des *big data* de prendre la place des chercheurs : d'après cet ouvrage, les chercheurs eux-mêmes *sont* les acteurs des *big data*. Édité par Microsoft Research, l'ouvrage est aux *data intensive-sciences* ce que serait pour la génomique le point de vue des fabricants de séquenceurs ADN : un point de vue certes légitime, celui des équipementiers, mais qui ne peut se substituer à la vision des chercheurs eux-mêmes et c'est d'ailleurs la critique majeure que l'on peut faire à l'ouvrage, car cette limitation n'est pas vraiment explicitée. Si la pratique scientifique est indubitablement vouée à évoluer en même temps que son instrumentation, sa visée en revanche demeure *a priori* inchangée : expliquer les phénomènes naturels. Ce n'est pas parce que l'équipement change que la finalité se voit transformée. Par ailleurs, il n'y a pas d'opposition incommensurable entre explication et prédiction, car expliquer, dans le paradigme galiléen, c'est dégager en langage mathématique les lois qui permettent de prédire le mouvement d'un corps par exemple, sa température, sa masse, etc. Qu'il existe alors désormais des outils computationnels pour manipuler ces lois, en déduire les conséquences, les calculer à partir de données, ne change pas ce paradigme dans lequel expliquer, c'est pouvoir prédire, mais le renforce au

contraire, en lui apportant davantage de moyens. Le but des sciences de la nature, qu'elles soient *data-intensive* ou théoriques, reste la production de connaissances explicatives.

En revanche, il convient de souligner que dans cette perspective instrumentaliste, les outils dits *big data* ne produisent pas les connaissances ; ils y contribuent seulement. Pour reprendre des exemples cités dans *The Fourth Paradigm*, le fait d'avoir à disposition des espaces de stockage informatique, un supercalculateur ou un logiciel de traitement d'image, ne suffit pas à produire de la connaissance, fût-ce avec une approche *data-intensive* : ils sont le prolongement du papier et du crayon, ou de la calculatrice, avec lesquels le mathématicien fait ses calculs. Les résultats doivent être contextualisés, interprétés, inscrits dans un protocole de recherche, pour avoir une signification scientifique.

Si l'on fait maintenant abstraction du contexte scientifique présenté par Hey, Tansley et Tolle, ces outils peuvent avoir d'autres usages, tout comme la calculatrice peut servir quasi-indifféremment au mathématicien, au comptable et au lycéen. Comme on l'a vu, l'informatique est un secteur économique prospère et les chercheurs ne sont pas les seuls à traiter de l'information sous format numérique, bien au contraire. Au niveau matériel (le *hardware)*, il n'y a quasiment pas de corrélation entre l'outil et l'usage. Les supercalculateurs sont généralement construits sur mesure pour un usage spécifique, mais l'énorme majorité des machines en fonctionnement, serveurs distants ou ordinateurs personnels, se prêtent tout aussi bien à une utilisation scientifique qu'à d'autres formes d'utilisation : c'est précisément leur polyvalence qui fait la force de nos ordinateurs actuels. Au niveau logiciel, la détermination de l'outil par l'usage est plus forte : outre les logiciels communs, comme les suites bureautiques, il y a une spécialisation forte en fonction de la tâche, même si des détournements sont toujours possibles. En revanche, les principes et procédés techniques qui sont à l'œuvre derrière ces différents logiciels spécialisés sont largement similaires, ce qui fait qu'un développeur est capable de contribuer à toutes sortes de logiciels au cours d'une carrière. Ce constat s'applique également aux informaticiens spécialisés en analyse computationnelle de données, ce qui fait que des domaines aussi variés que la reconnaissance vocale, la traduction automatique, les moteurs de recherche, la détection de pannes, reposent sur les mêmes techniques de traitement de l'information, manipulées par les mêmes personnes.

Cette distinction entre production de connaissances et traitement de l'information est cruciale, car elle fait partie des raisons pour lesquelles les acteurs des *big data* ne sont pas seulement les chercheurs et ceux qui travaillent pour eux. Outre le fait que toutes les connaissances ne sont pas produites par des chercheurs, il faut surtout retenir que le traitement de l'information à l'œuvre dans l'analyse computationnelle de données n'a pas systématiquement une visée scientifique, et cela est permis par la détermination seulement partielle de l'outil par l'usage. De fait, il existe tout un spectre de possibilités, entre une optique de représentation du réel par les données, et une optique d'intervention et de manipulation d'information. Ces deux extrémités se superposent partiellement à la distinction

qu'on a vue entre universités et entreprise, mais il n'est pas rare que d'autres combinaisons se présentent. Ainsi, pour reprendre l'exemple de Google, on y trouve chez cet acteur à la fois :

- des travaux scientifiques en traitement de l'information, visant à produire des connaissances et des savoir-faire dans ce domaine (par exemple ses contributions scientifiques en matière de *deep learning)* ;
- des travaux expérimentaux et prédictifs de représentation du réel, dont le domaine recouvre celui de l'épidémiologie (Flu Trends) ;
- et bien sûr, des outils opérationnels de traitement de l'information comme son moteur de recherche.

Il est difficile de placer ce dernier dans la dichotomie expliquer/prédire. Même si l'on peut suggérer que trier des résultats de recherche, c'est essayer de prédire l'information recherchée par l'utilisateur à partir des termes de sa recherche, cela revient à introduire dans la fonction d'un moteur de recherche un angle d'analyse qui ne s'y prête pas vraiment. Plus simplement, permettre d'accéder à de l'information n'est pas la même chose que de la produire, et a fortiori que de produire des connaissances explicatives ou prédictives. Plus largement, tous les outils du web qui reposent sur la recommandation de contenus (qu'il s'agisse de films, d'objets à vendre, ou de publications) et sont souvent cités en exemple d'application des *big data*, s'inscrivent, comme on le verra au chapitre VII, dans une visée qui n'est pas primairement d'expliquer ou de prédire. En revanche, dans des domaines comme l'actuariat, la finance ou le marketing, on rencontre des applications des *big data* dont l'objectif est de détecter un risque, une défaillance, une opportunité, et qui ont donc bien une visée prédictive. En synthèse, on dira que la distinction expliquer/prédire doit plutôt s'entendre comme un spectre, qui ne couvre pas, par ailleurs, la totalité des applications des *big data*.

3. La thèse de l'empirisme sans modèle

Une autre distinction récurrente dans la rhétorique des *big data* concerne l'opposition entre modèle et données, et notamment l'idée qu'avec suffisamment de données, les modèles deviennent inutiles car les données parlent d'elles-mêmes. Cette affirmation marque une distinction avec la statistique classique dont l'enjeu est précisément de construire et paramétrer des modèles pour ensuite les appliquer aux données dont on dispose. Un statisticien avec une formation classique sera non seulement en désaccord avec la thèse de l'obsolescence de la modélisation, mais il la dira impossible. Par exemple, Nate Silver, qui est statisticien, ne tient pas ce discours, bien qu'il soit l'un des champions des *big data*. Par ailleurs, la communauté scientifique est circonspecte quant à la capacité du « déluge de données » à produire des connaissances. Les mathématiciens Calude et Longo (2015) concluent ainsi un article sur les corrélations factices dans les *big data* :

> Anderson's recipe for analysis lacks the scientific rigour required to find meaningful insights that can change our decision making for the better. Data will never speak for itself, we give numbers their meaning, the Volume, Variety or Velocity of data cannot change that.

L'idée que les données puissent parler d'elles-mêmes est probablement l'une des plus polémiques et ambitieuses de la rhétorique des *big data* en tant que telle. C'est également une affirmation qui n'ignore pas l'approche classique des statistiques, et des mathématiques en général, mais la mobilise au contraire comme point de comparaison pour montrer qu'elle est radicalement meilleure. À titre d'exemple, la thèse de la supériorité des données sur les modèles a été défendue par Alon Halevy, Peter Norvig, and Fernando Pereira (2009), qui travaillaient alors tous les trois chez Google. Leur article « The Unreasonable Effectiveness of Data » est une référence explicite à un article du physicien Eugene Wigner, publié en 1960, « The Unreasonable Effectiveness of Mathematics in the Natural Sciences », qui illustre l'efficacité des formules mathématiques en sciences, et tout particulièrement en physique. Dans les termes de notre analyse, le texte de Wigner est une défense du paradigme galiléen, et de la prédilection des sciences de la nature pour un langage mathématisé. Nous lui devons cette célèbre citation :

> Le miracle de la justesse du langage des mathématiques pour la formulation des lois de la physique est un don merveilleux que nous ne comprenons ni ne méritons. [8]

D'après Halevy, Norvig et Pereira, ce miracle demeure en ce qui concerne effectivement la physique. Ils contestent en revanche l'extension de son domaine, et le fait qu'il puisse s'appliquer à d'autres sciences que la physique. En effet, écrivent-ils, « les sciences qui impliquent des êtres humains, plutôt que des particules élémentaires, se sont révélées plus réfractaires aux mathématiques élégantes ». Le domaine de connaissances visé par l'article est le langage, et plus spécifiquement le traitement automatique du langage naturel lorsqu'il mobilise des techniques d'apprentissage automatique *(machine learning).*

L'opposition formelle entre les « mathématiques élégantes » et l'informatique mérite également d'être discutée, ce que l'on fera au chapitre V. Disons provisoirement que d'un point de vue historique et philosophique, l'informatique théorique n'est rien d'autre que le prolongement des mathématiques. On rappellera notamment au chapitre V que l'article fondateur de Turing est une réponse au problème de la décision de Hilbert, un problème mathématique s'il en est. On pourrait dire que l'informatique théorique émerge comme une solution à un problème mathématique. Si l'on avance de quelques décennies, le *machine learning* n'est pas non plus radicalement opposé à l'approche statistique. Des travaux comme ceux de Vapnik (1998), sur la théorie de l'apprentissage statistique, unifient les deux approches. De nombreuses techniques de *machine learning* reposent sur des principes et outils

[8] "The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve."

statistiques, qu'ils soient fréquentistes ou bayésiens. Fondamentalement, l'enjeu du *machine learning* est de déduire un modèle à partir des données, soit, pour ainsi dire, d'automatiser le travail d'ajustement entre les données et le modèle qu'opère traditionnellement le statisticien.

Anderson est beaucoup plus radical dans son opposition entre modèle et données. Lorsque le statisticien George Box écrit, dans une formule célèbre, que tous les modèles sont faux, mais que certains sont utiles, il ne s'agit pas d'une invalidation externe de l'approche statistique en général, mais d'une critique interne de son imperfection. Anderson fait donc un contresens en le citant, ce qui l'amène à dire que les statistiques sont obsolètes. Dans le même esprit, mais avec un peu plus de modération dans le propos, Cukier et Mayer-Schönberger écrivent que la vraie révolution n'est pas dans les machines qui analysent les données, mais, dans la donnée elle-même et dans l'usage que l'on en fait[9] : ils ne se prononcent pas sur le travail de modélisation, non pas qu'il soit obsolète, mais qu'il n'est pas ce qui importe.

Pour résumer, on dira que d'un point de vue scientifique, il n'y a pas d'opposition fondamentale entre modèle et donnée ; ce sont des objets complémentaires tout comme un verre d'eau est constitué d'un récipient, le verre, et du liquide qu'il contient : ce sont bien évidemment des objets distincts, mais il ne nous appartient pas de les opposer comme s'ils étaient mutuellement exclusifs.

En réalité, la vraie opposition ne se situe pas entre donnée et modèle statistique. Comme nous le développerons au chapitre V, la confusion tient au fait que le terme « modèle » a un sens large, en sciences et en général. Pour simplifier, on dira provisoirement qu'il renvoie à trois types d'objets :

- les modèles scientifiques, qui sont des représentations qualitatives du réel ;
- les modèles mathématiques ou statistiques, qui expriment tout ou partie d'un modèle scientifique sous forme de systèmes d'équation ;
- les modèles informatiques, qui sont indépendants des précédents d'un point de vue théorique et portent sur la manière la plus optimale de calculer effectivement le modèle mathématique.

Pour bien comprendre Anderson (et la thèse de la prééminence de la donnée en général), il faut donc reformuler son opposition entre modèle et donnée comme une opposition entre modélisation rationnelle et empirisme. Dans celle-ci, on aurait d'un côté la théorie scientifique rationaliste, qui pose la raison humaine comme source et principe fondamental de la connaissance scientifique, et de l'autre l'empirisme, qui considère au contraire l'expérience comme la source des connaissances. En ces termes, l'opposition n'est plus d'ordre technique et statistique, mais d'ordre épistémologique, et prolonge dès lors une discussion qui traverse l'ensemble de l'histoire de la philosophie des sciences, de Platon au cercle de Vienne. La théorie et l'expérience étant deux composantes essentielles de la science actuelle, il n'est pas question de trancher en faveur de l'une ou l'autre de ces thèses : notre

[9] "the real revolution is not in the machines that calculate data but in data itself and how we use it"

position est que ces deux aspects ne sont pas exclusifs ou concurrents, mais complémentaires. D'un point de vue sociologique, il n'y a pas vraiment de prévalence ; à titre d'illustration, on peut citer une enquête publiée en 2013 par Bourget et Chalmers, et qui présente les réponses de près de 2000 philosophes (avec une surreprésentation de la philosophie anglo-saxonne et analytique) sur un certain nombre de débats philosophiques classiques. D'après cette enquête, 35% des répondants se disent empiristes, 27,8% rationalistes, et 37,2% ne se reconnaissent pas dans l'une ou l'autre de ces positions. Que les défenseurs des *big data* soient empiristes n'est donc pas intenable. En revanche il importe de caractériser la forme de l'empirisme qu'ils défendent, et la façon dont elle se distingue de formes antérieures.

La singularité de cet empirisme est son mode d'accès à l'expérience, si tant est que l'on s'accorde en effet à dire qu'il accède effectivement à l'expérience. Un empiriste comme Humphreys (2004) considère que les phénomènes naturels nous sont accessibles par observation directe, grâce à l'œil de l'observateur, ou indirecte, grâce à des instruments qui améliorent ou se substituent à la vision directe. Il y montre que l'on doit pouvoir superposer partiellement ce que voit l'œil et ce que montre un instrument, ou ce que montre un premier instrument, connu et considéré comme fiable, et un nouveau que l'on cherche à légitimer. C'est l'argument du recouvrement *(overlap)*, selon lequel il doit y avoir dans les modes d'accès une continuité par recouvrement. Dans « Est-ce qu'on voit à travers un microscope ? », Ian Hacking (1981) détaille la construction théorique nécessaire pour pouvoir dire que la réalité qui s'offre à nos regards, et celle que montre le microscope, sont bien les mêmes. Du point de vue de la perception, Daston et Galison (2007) ont montré combien le regard, y compris ce qu'il a de personnel et de subjectif, joue un rôle essentiel dans la constitution de l'objectivité scientifique. De manière générale, l'empirisme est un fondationnalisme qui prend un certain mode d'accès à l'expérience comme fondement apodictique de la connaissance scientifique (on peut penser à l'intuitionnisme de la phénoménologie husserlienne), et construit à partir de celui-ci une rationalité propre, qui explicite comment de nouveaux modes d'accès, notamment indirects, peuvent s'ajouter aux sources de la connaissance.

La forme d'empirisme défendue dans la rhétorique des *big data* pose de manière apodictique les données numériques comme source de la connaissance. De ce fait, elle ne s'inscrit pas dans le fondationnalisme de l'empirisme classique, étant donné qu'elle n'explicite pas comment relier la perception directe ou indirecte et les données numériques. De ce point de vue il y a une rupture nette avec les sciences classiques : l'empiricité de la donnée numérique n'est pas justifiée à l'intérieur du cadre des sciences classiques, mais posée en tant que telle comme un nouveau cadre, avec son propre mode d'accès au réel, ses outils, ses méthodes, ses principes. Elle apparaît davantage comme un *ethos*, c'est-à-dire une norme ou une convention sur la manière de faire, que comme quelque chose qui ressort nécessairement des pratiques. D'un point de vue ontologique, c'est la réalité elle-même qui en est transformée : les données numériques ne sont pas une représentation du réel, mais bien le réel lui-

même. Ce point est à la fois l'un des plus contestables, mais aussi l'un des plus fondamentaux et caractéristiques de la promesse des *big data*. Il faut l'accepter ou le refuser mais il se pose comme un axiome indiscutable. La promesse des *big data* est celle d'un empirisme qui remplace l'observation et la mesure par des données numériques, c'est-à-dire un réel toujours déjà manipulable et calculable. Pour paraphraser le journaliste Michel Mourlet parlant du cinéma, les *big data* sont un regard qui se substitue au nôtre pour nous donner un monde accordé à nos outils de manipulation computationnelle[10].

Il est vrai que dans les *data-intensive sciences*, comme par exemple en génomique, au moins une partie des données utilisées sont bien considérées comme des mesures des phénomènes naturels, et il n'y a pas cette rupture entre l'observation et la donnée numérique, qui n'est que le format de la mesure. La représentation informatique de ces données, leurs modalités de stockage, posent quelques difficultés où l'épistémique s'immisce dans une discussion qui peut sembler purement technique, mais dans l'ensemble des solutions de continuités sont trouvées pour donner sens à ces données dans le cadre du modèle ou paradigme scientifique à l'œuvre. Dans ce contexte, c'est le traitement réservé aux données qui reconfigure leur empiricité, ainsi que la possibilité de mêler des données classiques issues des instruments scientifiques avec d'autres sources : ainsi, en santé peuvent se mêler les données médicales « classiques » et les données produites par les assureurs, les objets connectés de mesure de soi, des données collectées sur les réseaux sociaux, etc.

Au cours des traitements qui leurs sont appliqués, toutes ces données sont manipulées uniquement en tant qu'elles-mêmes, *comme si* elles ne provenaient pas de mesures directes ou indirectes des phénomènes. Ainsi, la bio-informatique utilise massivement des techniques provenant d'autres domaines, comme par exemple les modèles de Markov cachés ou les machines à vecteur de support, au prix d'une généralisation puis d'une re-spécialisation de ces techniques. Quelle que soit la provenance des données, qu'elles soient nativement numériques ou numérisées, les traitements qui y sont appliqués reposent sur une cohérence et des normes épistémiques métissées, qui mêlent d'un côté les contraintes posées par les techniques computationnelles (et les modèles informatiques et statistiques sur lesquels elles reposent), et d'autre part les concepts scientifiques qui y sont injectés pour que les résultats soient interprétables dans un certain cadre théorique. Dans ce contexte, l'hétérogénéité des normes qui régissent les techniques utilisées est, comme on le verra au chapitre V, ce qui confère aux *data-intensive sciences* leur caractère « *big data* ».

Qu'il s'agisse donc des *data-intensive sciences* ou des *big data* en général, ce que l'on formulait initialement comme une opposition entre modèle et donnée peut ainsi être retraduit dans les termes suivants :

[10] Incorrectement cité dans *Le Mépris* par Jean-Luc Godard qui attribue la citation à André Bazin, Michel Mourlet écrivait dans son article « Sur un art ignoré » : « Le cinéma est un regard qui se substitue au nôtre pour nous donner un monde accordé à nos désirs ».

- de nouveaux modes de construction des modèles statistiques et scientifiques ;
- une prédilection pour un empirisme qui pose la valeur apodictique de la donnée, au détriment du rationalisme ;
- une hétérogénéité (sur laquelle nous reviendrons au chapitre V) entre les modèles eux-mêmes d'une part, et entre les normes de constitution des modèles et les normes de recueil des données d'autre part.

Ces normes de recueil de données ont un impact tout particulier sur la nature des données collectées et les régimes de constitution d'ensemble (ou jeux) de données.

4. La thèse de l'exhaustivité des données

L'une des affirmations les plus marquantes que l'on retrouve tout particulièrement chez Cukier et Mayer-Schönberger est l'idée selon laquelle le développement des *big data* rend obsolète la pratique de l'échantillonnage statistique et donne accès, sans biais, à la totalité des données, ce qu'ils résument par « N=all ». Ils écrivent ainsi :

> One of the areas that is being most dramatically shaken up by N=all is the social sciences. They have lost their monopoly on making sense of empirical social data, as big-data analysis replaces the highly skilled survey specialists of the past. The social science disciplines largely relied on sampling studies and questionnaires. But when the data is collected passively while people do what they normally do anyway, the old biases associated with sampling and questionnaires disappear. We can now collect information that we couldn't before, be it relationships revealed via mobile phone calls or sentiments unveiled through tweets. More important, the need to sample disappears.

Au-delà des accents millénaristes avec lesquels ils expriment leur position, la position de Cukier et Mayer-Schönberger s'inscrit sans le savoir dans un débat épistémologique plus ancien que la période historique, relativement récente en réalité, qu'ils décrivent au travers des enquêtes et sondages. L'histoire de l'articulation entre statistique et société a notamment été étudiée par Alain Desrosières (2010), mais aussi, entre autres, par Cohen (2005), Bowker et Star (1999), Stigler (1999), Porter (1995), Gigerenzer et al. (1989). Ces travaux représentent aujourd'hui un domaine de recherche assez clairement constitué qui interroge à la fois les normes et les méthodes statistiques d'un point de vue plutôt internaliste, et leurs relations avec les individus, la société et le pouvoir dans une perspective plus externaliste d'histoire et de sociologie de la quantification. Ils permettent de réinscrire la pratique du sondage aléatoire dans un temps historique plus long. Ainsi, jusqu'au XXe siècle, ce qui ne s'appelle pas encore les sciences sociales, mais l'arithmétique morale ou politique, s'efforce, avec plus ou moins d'optimisme et de naïveté, d'atteindre l'exhaustivité statistique par le recensement : dans cette perspective, les *big data* ne seraient pas tant une révolution scientifique qu'un retour aux sources méthodologique (ou une régression, selon l'appréciation que l'on peut porter à cette évolution).

Les travaux d'Alain Desrosières ont montré combien l'ambition d'exhaustivité était difficile à réaliser, et comment la pratique de l'échantillonnage aléatoire s'est développée comme une façon d'éviter les biais inhérents à des recensements certes massifs, mais incomplets. La prétention à l'exhaustivité a été largement critiquée dans son fondement même, non pas pour dire qu'elle n'est pas épistémologiquement souhaitable, mais parce qu'elle est inaccessible d'un point de vue théorique et pratique (Bowker, 2014; boyd & Crawford, 2011; Lagoze, 2014; Leonelli, 2014). On peut tendre vers l'exhaustivité, comme un horizon, mais on ne l'atteint jamais. Comme le résume Carl Lagoze (2014) :

> [...] any data, no matter what its size, is de facto a sample, with bias implicit due to choice of instrumentation, span of observation, units of measurement, and numerous other factors. In essence, *n* never equals all; all is a limit in mathematical terms that can be approached but never attained.

Comme le modèle physique, l'échantillon est une représentation simplifiée de son objet (la population), avec lequel il partage un certain nombre de propriétés jugées pertinentes. Comme on le verra au chapitre IV, cette représentation est construite soit par tirage aléatoire (c'est la méthode la plus fiable pour les statisticiens), soit par ajustement de quotas dont les proportions au sein de la population sont connues (c'est la méthode la plus répandue, car plus facile à mettre en œuvre, notamment dans les instituts d'étude). Cet échantillon de population est soumis à un certain nombre de questions visant à recueillir leur opinion. Aux erreurs de mesure près, cette opinion est considérée comme une image (ou une icône, dirait Peirce) de l'opinion de la population. Les biais de cette méthodologie introduits par les problèmes de sélection de l'échantillon, la formulation du questionnaire, les modalités d'administration du questionnaire ou encore le taux de non-réponse par type de sous-population, sont difficilement mesurables ponctuellement, mais généralement connus, documentés, et considérés comme acceptables. Que l'on aborde le succès de cette méthodologie sous un angle internaliste, en s'intéressant au succès des résultats obtenus, ou sous un angle externaliste, comme le fait Desrosières en parlant de routinisation et de stabilisation de ces méthodes, il faut à tout le moins admettre que l'échantillonnage constitue aujourd'hui la norme des sciences sociales quantitatives.

De ce fait, proposer une méthodologie alternative dans un cadre continuiste impliquerait en toute logique de discuter ces différents points et de montrer comment la nouvelle méthode proposée permet de préserver leurs avantages, ou d'en présenter de meilleurs. La continuité entre les méthodes classiques et l'introduction de « *big data* » est d'ailleurs le point de vue défendu par Sabina Leonelli (2014) en ce qui concerne la biologie :

> [...] data quantity can indeed be said to make a difference to biology, but in ways that are not as revolutionary as many *Big data* advocates would advocate. There is strong continuity with practices of large data collection and assemblage conducted since the early modern period; and the core methods and epistemic problems of biological research, including exploratory experimentation, sampling and the search for causal mechanisms, remain crucial parts of inquiry in this area of science [...]

De manière similaire, Carl Lagoze (2014) analyse l'arrivée de grands volumes de données dans le cadre des pratiques existantes et s'efforce de mettre en évidence une distinction entre « beaucoup de données » et les « *big data* » à proprement parler. Dans le premier cas, on observe une augmentation de la quantité de données essentiellement quantitative, qui soulève des enjeux techniques et méthodologiques, mais est traitée avec des solutions de continuité par rapport au cadre épistémologique en place. Il s'agit notamment de contextualiser et documenter les données, notamment lorsqu'elles circulent d'un chercheur à l'autre, pour expliciter leur signification et la façon dont elles peuvent être analysées. Dans le second cas, le changement est de nature qualitative, et remet en cause le cadre scientifique ; il est en rupture avec le paradigme existant, avec des conséquences principalement négatives. Du point de vue de l'épistémologie classique, cette rupture induit en effet une perte de contrôle épistémique et de confiance dans l'intégrité des données, ce qui n'est pas acceptable pour les sciences traditionnelles. Au prisme de l'existant, les *big data* ne représentant pas tant un progrès qu'une crise de la production de connaissances.

Aborder les *big data* dans un cadre continuiste n'est donc pas une bonne manière de convaincre de la valeur épistémique qu'elles peuvent apporter. Ce n'est d'ailleurs pas l'approche de Cukier et Mayer-Schönberger puisqu'ils s'inscrivent, comme on l'a vu, dans une rhétorique de la rupture ; il est d'ailleurs plutôt surprenant que les sciences sociales et l'échantillonnage soient même mentionnés. Ce qu'il faut cependant souligner, comme dans la question de l'empirisme, c'est qu'il ne s'agit pas d'un simple retour aux sources, mais, comme on le verra au chapitre III, d'une nouvelle forme de recherche d'exhaustivité qui présente deux différences majeures par rapport au recensement :

- l'enregistrement de l'information est systématique, effectué par défaut, par le dispositif lui-même, ce qui suggère une plus grande couverture (mais non une exhaustivité, car il existe des enjeux techniques et juridiques de perte ou de suppression de données) ;
- l'enregistrement se fait au niveau de l'outil, et l'outil n'est pas l'individu : les détenteurs de téléphones portables, les utilisateurs de Twitter, sont des populations particulières du point de vue sociodémographique qui intéresse le recensement classique.

L'exhaustivité de l'information numérique n'est donc pas l'exhaustivité de l'activité sociale et culturelle. Elle abolit cependant le principe de la représentativité comme mode de validation des données. Devant cet état de fait, on rencontre, comme on le verra, plusieurs attitudes, telles que la tentative de réinstaurer une forme classique de représentativité par redressement statistique, la recherche d'autres formes de représentativité en rupture avec celle des études d'opinion, ou encore l'absence de discussion sur la question particulière de la validation des données, celles-ci étant considérées en tant qu'elles-même.

Par ailleurs, cette exhaustivité de la donnée numérique est une exhaustivité théorique, bordée d'un côté par des limitations socio-économiques et de l'autre par des limitations techniques. D'un point de

vue économique et social en effet, tous les acteurs n'ont pas le même pouvoir d'accès à l'information. Les détenteurs de données (dont notamment les réseaux sociaux majeurs) disposent d'un accès privilégié à l'information (Manovich 2011) et ont instauré la mise en visibilité des données comme une stratégie à part entière. Les agents collecteurs de données n'ont pas accès à tout, et ne savent pas forcément ce qu'ils ne voient pas, à quoi ils n'ont pas accès. Par exemple, Twitter ne donne accès qu'à des recherches précises (et limitées en volume) ou à un échantillon de l'ensemble des tweets produits. Qui plus est, cet échantillon ne serait pas représentatif de l'ensemble des tweets produits du point de vue du volume, de la fréquence des mots, ou encore de la structure des interactions qui s'en dégage (Morstatter et al. 2013). Google proposait jadis un accès à son moteur de recherche sous forme d'API, mais a fermé le service en 2009, suivi par Yahoo! puis Bing, qui en fournit désormais une version payante (Tanguy 2013), mais réputée moins complète que ce que Google pourrait proposer. Du fait du rôle crucial joué par les moteurs de recherche en matière d'accès à l'information web, et du rôle techniquement et économiquement dominant de Google dans cette tâche, les fournisseurs de logiciel de veille et de recherche d'information déploient des stratégies de contournement de cette limitation. Pour cela, ils s'efforcent de faire passer leurs crawlers, les robots de collecte d'information, pour des humains consultant le moteur de recherche, par exemple en ajoutant des temps de pause aléatoire entre les actions de défilement et de clic sur une page. De ce fait, c'est une véritable course à l'armement qui se joue entre ce type d'acteurs et les détenteurs de données, qui développent des stratégies sophistiquées de détection des comportements automatisés, matérialisées par exemple par la nécessité de compléter un captcha pour pouvoir continuer à consulter des pages de recherche de Google. Dans une autre configuration, Facebook ne donne accès qu'à l'information publique, et à l'information privée visible de l'utilisateur collecteur de données, telles que les actions et publications de ses contacts, éventuellement celles des contacts de ses contacts : au lieu de renvoyer une certaine image de la population des internautes, Facebook renvoie à l'agent une image de lui-même et de son réseau social. Une stratégie de contournement de cette limitation par la collaboration a été expérimentée par le projet ANR Algopol (Bastard et al. 2013), dans le cadre duquel une application intégrée à Facebook proposait aux utilisateurs bénévoles de donner aux chercheurs un accès à leurs données personnelles. À partir de ces données, un échantillon représentatif de la population a été constitué par quota (en partenariat avec un institut de sondage), dans l'espoir de pouvoir généraliser les observations sur ces données personnelles et les modes d'utilisation de Facebook qu'elles reflètent à la population française (ou en toute logique, à la population française utilisatrice de Facebook).

Comme on le reverra dans les chapitres suivants, l'accès à l'information est donc également régi d'un point de vue technique : l'agent qui sait se connecter à une API, contourner les protections mises en place par Google, développer une application spécifique, aura, en qualité comme en quantité, des données différentes d'un agent moins compétent techniquement. Par ailleurs, la technicité de l'informatique, qui s'exprime dans les opérations de collecte de données, est aussi le règne de l'erreur,

de l'accident. Les problèmes de format, de transmission d'information, d'incompatibilité entre agents logiciels, etc., sont le quotidien de l'agent collecteur de données. À titre d'exemple, Twitter attribue normalement à chaque tweet un identifiant unique et définitif, constitué d'une série de chiffres ; en pratique, nous avons pu constater que certains tweets changeaient d'identifiant ou qu'un même identifiant avait pu être attribué à des tweets différents. Les outils de collecte de données se construisent suivant certaines hypothèses qui peuvent être temporairement ou définitivement invalidées, du fait des modifications stratégiques et techniques opérées par les détenteurs de données, tout comme de la complexité des systèmes informatiques sujets à des pannes (absence de fonctionnement) et des dysfonctionnements (fonctionnement anormal). Par ailleurs, il est bon de souligner que les supports et formats numériques ne sont pas atemporels. La stabilité des dispositifs informatifs dépend de modifications constantes. Comme pour l'aristocratie italienne de Lampedusa, il faut continuellement tout y changer pour que rien ne change. Ces dispositifs s'appuient sur des langages et des protocoles qui n'ont de standard que le nom, dans la mesure où plusieurs standards coexistent, et que les tentatives d'unification aboutissent généralement à la création d'un nouveau standard concurrent des précédents ; ces standard changent, apparaissent et disparaissent. Pour des motifs de sécurité ou de performance par exemple, la mise en visibilité proposée par le détenteur des données peut changer, rendant incompatibles et inopérants les dispositifs de collecte qui en dépendent : tous ces dispositifs doivent être appréhendés dans une ontologie de la fragilité, une matérialité qui ne persiste dans le temps que par la maintenance constante des « petites mains » (Denis et Pontille 2011). Ce besoin d'attention constant peut conduire les collecteurs de données à certains choix techniques nécessitant un effort moindre, au prix de la disparition de certaines données ou métadonnées.

Du fait de toutes ces limitations socio-économiques et techniques, l'agent qui constitue un jeu de données aura beaucoup de mal à le caractériser d'un point de vue statistique : les données manquantes, mal formatées, inaccessibles pour lui, constituent non seulement un silence statistique par rapport à l'information recherchée, mais un silence dont il n'a pas forcément connaissance, pour peu, par exemple, qu'il n'ait pas développé lui-même, ou pas entièrement, l'outil de collecte qu'il utilise. De ce fait, la qualification statistique des données collectées, dont on a vu qu'elle était difficile en théorie, peut s'avérer tout simplement impossible en pratique. Au lieu de l'exhaustivité annoncée, on découvre ainsi une quantité peut-être vertigineuse, mais inqualifiable et empreinte d'incertitudes. Cette quantité, enfin, ne dit rien par elle-même du fait humain, mais désigne seulement un fait numérique, une accumulation des enregistrements d'actions sur des supports informatisés.

5. La thèse de l'abandon de la recherche des causes

De cette quantité sans signification émergent cependant des régularités, des tendances, des grandes masses. Dans le projet Google Flu Trends comme dans d'autres, on ne recherche plus les causes du

phénomène (en l'occurrence l'épidémie de grippe) mais la détection de son existence, que l'on cherche à appréhender de manière quantitative : à travers son ampleur, sa progression. L'épistémologie des *big data* entend abolir la notion de causalité pour la remplacer par celle, désormais suffisante, de corrélation. Cukier et Mayer-Schönberger écrivent ainsi :

> The era of *big data* challenges the way we live and interact with the world. Most strikingly, society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what. This overturns centuries of established practices and challenges our most basic understanding of how to make decisions and comprehend reality.

Dans le même esprit, Anderson déclare :

> There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

En deux courtes citations, plusieurs conceptions de la causalité se dégagent :

- comme réponse à la question « pourquoi » ;
- comme manière de comprendre la réalité ;
- comme forme de modèle ou d'hypothèse ;
- comme caractérisation de la science.

Nous allons voir que chacune de ces conceptions est inexacte à plusieurs égards. Ainsi, on considère traditionnellement qu'une réponse à une question commençant par « pourquoi » est une explication, à laquelle on peut répondre de plusieurs façons. Dans la théorie aristotélicienne de la causalité, on distingue quatre causes : la cause matérielle (la matière qui constitue la chose), la cause formelle (l'essence de la chose), la cause motrice (ce qui l'a constituée) et la cause finale (la finalité de la chose). Cette théorie de la causalité renvoie à une métaphysique téléologique où chaque chose existe en vue d'une finalité. D'un point de vue épistémologique, il s'agit d'un idéalisme qui fait de la science la connaissance des causes par la démonstration et indépendamment de l'expérience. En ces termes, les *big data* s'opposent effectivement à la conception aristotélicienne des sciences et de la causalité.

Cependant, cette conception est loin d'être la plus en vogue en science et en philosophie contemporaines, pour ne pas la dire obsolète. À partir du XVIIe siècle, la science des modernes s'appuie massivement sur la notion de déterminisme mécanique, et sur l'idée que les propositions scientifiques expliquent, à partir des lois de la nature et de conditions initiales, *comment* les phénomènes naturels se produisent. La question de savoir si la causalité porte sur le *pourquoi* ou le *comment* est reléguée au statut de question métaphysique, dont la réponse ne peut être déterminée ni par les sciences modernes, ni par les *big data*. Or, les corrélations telles que celles qu'on observe dans les données numériques ne peuvent exister sans une forme de déterminisme qui les rend régulières,

prévisibles, indépendamment du fait que l'on cherche à en formuler une explication téléologique (comme chez Aristote) ou nomologique (comme dans les sciences de la nature modernes). De ce point de vue-là, les *big data* représenteraient la fin de la quête du *comment* (et non du *pourquoi)* pratiqué par les sciences de la nature moderne, tout en restant dans un cadre déterministe.

Il n'est pas certain, cependant, que l'abolition du *comment* soit possible. La philosophie moderne de la causalité s'appuie notamment sur Hume, chez qui la causalité n'est pas tant une propriété métaphysique des choses qu'une nécessité de notre esprit pour appréhender le monde et donner une continuité à nos perceptions singulières. Il définit ainsi la relation entre une cause et un effet comme une contiguïté spatio-temporelle, où la cause précède l'effet et où celui-ci la suit avec régularité, ou autrement dit, s'inscrit dans une relation nécessaire avec elle. Chez Hume, la recherche des causes n'est pas tant le but de la science qu'une disposition de l'esprit humain, une habitude du sens commun dont on ne peut se défaire ; empiriquement, on n'observe rien d'autre que des cooccurrences qui sont interprétées comme des relations de cause à effet. Dans cette perspective, les *big data* auront beau ne présenter que des régularités sans signification, notre esprit sera toujours naturellement enclin à les interpréter comme des relations de cause à effet.

Au XXe siècle, des philosophes plus contemporains de la causalité tels que Reichenbach, Suppes ou Cartwright examinent plus précisément les difficultés posées par cette interprétation, et l'insuffisance du cadre formel de l'inférence causale pour distinguer les relations causales des corrélations trompeuses. En lien avec le développement du domaine, ces auteurs s'intéressent plus particulièrement à la causalité statistique. Dans son livre *Causes, probabilités, inférences*, Isabelle Drouet (2012) examine les rapports entre ces théories de la causalité et les inférences causales pratiquées en probabilité et en statistique. L'étude de ces inférences nous intéresse tout particulièrement car les outils statistiques examinés, tels que les réseaux bayésiens, sont à la base de nombreuses techniques utilisées dans les *big data* pour faire émerger des corrélations.

Isabelle Drouet souligne que la causalité statistique est une causalité générique, par opposition à une causalité singulière qui concerne un événement singulier (par exemple, on distingue d'un côté « fumer cause le cancer du poumon » et de l'autre « le tabagisme de Pierre a causé son cancer »). Le cadre de cette causalité générique n'est plus mécanique mais statistique : les inférences statistiques causales ne cherchent pas à faire émerger des relations nécessaires entre une cause unique et son effet, mais des relations probables, des facteurs qui peuvent être multiples et qui font augmenter la probabilité d'un effet au niveau statistique, sans que la relation ne soit forcément systématique. Ces inférences sont hypothético-déductives mais peuvent aussi être inductives, et ne s'appuient donc pas nécessairement sur des hypothèses préalables.

Dans son ouvrage, Isabelle Drouet rappelle également que les théories contemporaines de la causalité ne sont pas unifiées, et souligne que les philosophes contemporains de la causalité défendent plutôt

une forme de « pluralisme causal » qui admet une complémentarité entre les différentes théories de la causalité. De fait, les défenseurs des *big data* ne rentrent pas dans cette discussion en explicitant quelle théorie de la causalité ils entendent rejeter. Cependant, il est plus difficile de réfuter une pluralité de points de vue sur la causalité qu'un réalisme dogmatique quant à son essence.

De cet ouvrage ressort également une réduction des ambitions de la définition de la causalité. Devenue conditionnelle, factorielle, la causalité n'a plus le statut de raison nécessaire et suffisante qu'on pouvait lui trouver chez les modernes, comme chez Descartes. La multiplication des conditions pour qu'une relation puisse être interprétée comme causale n'enlève rien au fait qu'il s'agit d'un cadre formel, et donc insuffisant pour assurer qu'on puisse effectivement identifier une relation de causalité. Dans l'analyse des facteurs d'une maladie par exemple, une analyse probabiliste ne pourra être concluante que si l'on a bien identifié et mesuré tous les facteurs candidats à favoriser le développement de la maladie, et qu'ils figurent dans les données sur lesquelles porte l'analyse. Sans cela, l'analyse ne pourra faire émerger que des causes mineures ou factices.

L'existence de corrélations accidentelles, de causes communes et inconnues ayant plusieurs effets, corrélés entre eux de ce fait (ce que Reichenbach appelle une « fourche conjonctive »), d'effets antérieurs à leurs causes, de facteurs qui s'annulent, mettent en difficultés des théories de plus en plus raffinées et d'une technicité croissante. Les praticiens des *big data* Fan, Han et Liu (2013) soulignent d'ailleurs que l'augmentation du volume de données analysées accroît le risque d'apparition de corrélations factices, corrélées de manière accidentelle entre elles (endogénéité) ou avec le bruit statistique produit par la diminution de la qualité des données. Le « déluge de données » engendrerait un « déluge de corrélations factices » (Calude et Longo 2015). La baisse de qualité des données est elle-même liée au moindre soin apporté à la constitution du jeu de données (on privilégie la quantité à la qualité) et à la multiplicité des sources de données, qui peut entraîner une augmentation des biais (c'est-à-dire des erreurs systématiques) présents dans les données.

Néanmoins, malgré les raffinements comme ceux de Reichenbach, Isabelle Drouet conclut qu'il

> n'est pas possible de s'appuyer sur les théories probabilistes de la causalité afin d'induire des connaissances relatives aux relations de cause à effet.

En d'autres termes, il y aurait, dans une perspective réaliste, de « vraies » relations de cause à effet, que les théories probabilistes de la causalité ne parviendraient pas à capturer. Par ailleurs, les outils et méthodes statistiques d'inférence causale ne respecteraient que grossièrement ces théories probabilistes de la causalité, et ne seraient pas fiables pour autant.

Si les critiques (au sens de limites) formulées par Isabelle Drouet envers ces théories sont plutôt de nature logique et formelle, nous aimerions y adjoindre des réserves de nature plus pragmatique. En effet, les outils statistiques matérialisent des modèles et des représentations de la causalité, sans jamais

produire des résultats de nature effectivement causale. Comme le souligne Hume en parlant des habitudes du sens commun, et donc du statut épistémologique plutôt qu'ontologique de la causalité, ces outils ne font jamais apparaître que des corrélations statistiquement significatives, qui sont ensuite *interprétées* comme des relations causales. D'un point de vue purement probabiliste, une relation reçoit une *probabilité quantifiée* d'être de nature causale, et qui est ensuite infirmée ou confirmée. Cette probabilité elle-même est sujette à caution : elle peut ne pas être fiable, et dépend des données sur lesquelles elle a été calculée. Les outils statistiques produisent des corrélations, et les individus qui les évaluent formulent des relations de causalité.

En mobilisant plusieurs théories de la causalité, d'Aristote aux réseaux bayésiens, nous avons établi que l'analyse causale cherche à répondre à la question « comment » plutôt qu'à la question « pourquoi », qu'elle relève de l'explication et non de la compréhension, qu'elle n'est pas l'apanage des sciences et n'est que faiblement liée à une approche hypothético-déductive. Sous ces aspects, elle n'apparaît pas comme une notion particulièrement pertinente pour démarquer les *big data* des sciences traditionnelles. Dans ces deux cadres, les outils statistiques font apparaître des corrélations que l'humain peut interpréter, à tort ou à raison, comme des relations de causalité.

Une démarcation plus raffinée consisterait à examiner précisément les formes de corrélation à l'œuvre dans les *big data*. En effet, si l'explication scientifique s'intéresse plus particulièrement aux corrélations qui peuvent être interprétées comme des relations causales, une analyse purement statistique (qu'elle soit destinée ou non à un usage scientifique) peut également mettre en évidence des relations purement corrélatives, sans lien causal, et où il n'y aurait donc pas de sens à chercher un lien de cette nature. Par exemple, deux caractéristiques d'un individu, comme l'âge et le statut marital, peuvent être régulièrement corrélées sans qu'il y ait de rapport causal entre ces propriétés. Un individu a une probabilité plus élevée d'être célibataire en-dessous d'un certain âge, marié puis veuf à partir d'un certain âge. Cela ne veut pas dire que l'âge est la cause du célibat, du mariage ou du veuvage.[11] Dans un autre contexte, l'analyse des centres d'intérêt, des pages consultées, des interactions, à l'œuvre notamment dans le ciblage publicitaire, repose sur l'identification de ces corrélations qui vont permettre à l'une des variables de servir d'intermédiaire ou de substitut à l'autre en cas d'information manquante. Un internaute qui consulte des magazines féminins en ligne a plus de chances d'être une femme qu'un homme ; un internaute qui consulte des médias couvrant l'actualité technologique est vraisemblablement intéressé par ces sujets. Une régie publicitaire qui a accès à cet historique de recherche pourra ainsi leur proposer des publicités pour des produits destinés aux femmes ou aux technophiles. Des chercheurs en sciences sociales de l'Université de Cambridge

[11] Evidemment, on peut considérer que l'expérience subjective du vieillissement est une motivation pour se marier, et que l'âge du conjoint, souvent corrélé à celui de l'individu, fait augmenter la probabilité du décès du conjoint et donc de l'individu. On rencontre aussi des individus pour lesquel le vieillissement sera une motivation pour rester célibataire, de jeunes veufs et des personnes âgées qui ne seront jamais veuves. Néanmoins, on est ici précisément dans le type de relation potentiellement causale qui demande une interprétation et ne se déduit pas mécaniquement d'une analyse statistique.

et de Microsoft Research (Kosinski, Stillwell et Graepel 2013) ont mis en évidence des corrélations autour de 90% entre les centres d'intérêts déclarés par les utilisateurs de Facebook (les « Like ») et des variables comme l'origine ethnique, l'orientation sexuelle ou politique (telles que renseignées *via* un questionnaire par un panel d'utilisateurs de Facebook). Ils écrivent par exemple :

> Good predictors of male homosexuality included "No H8 Campaign," "Mac Cosmetics," and "Wicked The Musical," whereas strong predictors of male heterosexuality included "Wu-Tang Clan," "Shaq," and "Being Confused After Waking Up From Naps."

Mettre en évidence des indicateurs statistiquement significatifs de l'orientation sexuelle ne permet pas d'affirmer qu'il y a une relation causale entre, par exemple, le fait d'aimer les comédies musicales et l'orientation sexuelle, dans un sens comme dans l'autre. Un chercheur en sciences sociales, du moins, ne fera pas d'affirmation de cette nature sans une investigation complémentaire, une conceptualisation et une analyse des dynamiques sociales et culturelles liées, dans notre exemple, à l'orientation sexuelle. Il n'est pas certain qu'il soit possible d'exprimer une relation causale de quelque nature que ce soit entre l'orientation sexuelle et certains centres d'intérêt, mais il y a bien une corrélation entre les deux, si l'on admet que l'écart entre les centres d'intérêt déclarés par la personne sur Facebook et ses centres d'intérêts « réel » (ou recueillis par des méthodes plus classiques) est suffisamment faible pour qu'on puisse les assimiler. Cette corrélation est suffisamment significative d'un point de vue statistique pour que la variable « centre d'intérêt » puisse être utilisée comme substitut *(proxy)* de la variable « orientation sexuelle ». Nous reviendrons, au chapitre suivant notamment, sur la nature et les modalités de ces substituts.

Dans les *big data*, il ne s'agit donc pas tant de supprimer l'interprétation causale des relations entre variables que d'étudier d'autres formes non causales de relations, de manière à pouvoir reconstituer des informations manquantes sur les individus par l'intermédiaire de variables substitutives.

6. La thèse de l'inductivisme

La question de l'empirisme et du corrélationnisme conduit assez naturellement à la question des types de raisonnements épistémiques[12], ou autrement dit la question de savoir si les raisonnements à l'œuvre dans les *big data* sont plutôt de nature inductive ou déductive. Pour Domingos, la réponse est sans ambiguïté en faveur de l'inductivisme auquel il consacre un chapitre assez curieux d'un point de vue philosophique, où il proclame Hume plus grand philosophe de tous les temps, et propose une réécriture surprenante du pari pascalien où Dieu est remplacé par un algorithme de *machine learning*. Si l'on prend d'un côté l'association empirisme-corrélation-induction, et de l'autre l'association rationalisme-causalité-déduction, les défenseurs des *big data* soutiennent systématiquement la

[12] Nous utilisons ici « épistémique » plutôt que scientifique, comme c'est l'usage, du fait que le critère de démarcation entre scientifique et non scientifique (ou pseudo-scientifique) s'est révélé, comme on l'a vu, non pertinent pour notre analyse.

première. Néanmoins, si l'on regarde de plus près les raisonnements à l'œuvre dans les pratiques effectives d'analyse computationnelle de données, et les réflexions épistémologiques qui en ressortent, on découvre une discussion bien plus complexe. En statistique notamment, on rencontre en effet des inductivistes et des déductivistes, mais aussi des positions qui tendent à montrer que les méthodes statistiques sont à la fois inductivistes et déductivistes, ou que cette opposition n'est pas forcément pertinente pour caractériser les statistiques en tant que telles. La statistique classique (aussi appelée « fréquentiste ») de Pearson et Fisher est souvent vue comme l'instrumentum des raisonnements scientifiques hypothético-déductifs, tandis que les statistiques bayésiennes seraient d'obédience inductiviste. Jean-Paul Benzécri, fondateur de l'école française de statistique, qui est souvent vue comme nettement inductiviste, présente Fisher comme un inductiviste. Tout en soulignant cependant que

> Le statisticien qui dans l'abondance des faits individuels découvre des lignes, globales, prétend faire œuvre inductive : aussi est-il porté à magnifier l'induction. (Benzécri 1976b)

il rappelle également qu'

> induction et déduction ne peuvent être séparées; il faut donc se garder de les opposer.

Dans le même esprit d'atténuation de l'opposition entre induction et déduction, Gelman et Shalizi (2013) s'opposent à une philosophie inductive pour analyser les statistiques bayésiennes et suggèrent que toutes les écoles statistiques permettent des inférences inductives dans un cadre qui est cependant généralement déductiviste, en ce qui concerne l'utilisation scientifique des statistiques notamment. Isabelle Drouet (2012) montre que les réseaux bayésiens peuvent tout aussi bien se prêter à un raisonnement hypothético-déductif qu'à des inférences inductives. Dans l'ensemble, la statistique inférentielle apparaît comme un outil au service de raisonnements qui peuvent tout aussi bien être inductivistes que déductivistes, voire favoriser le premier à un niveau plus local, tout en mobilisant le dernier à un niveau plus global, et réciproquement. Les auteurs que nous avons cités contestent tous l'idée d'une détermination nécessaire, par son *instrumentum*, de la forme de raisonnement utilisé.

Les analyses computationnelles à l'œuvre dans les *big data* héritent pour une bonne part de cette configuration, pour le moins lorsque les acteurs en question sont des statisticiens. Les informaticiens (comme Domingos) se présentent plutôt comme des inductivistes, même s'ils sont bien souvent les premiers à mobiliser dans leurs raisonnements des assertions d'ordre général et posées axiomatiquement. Par exemple, dans un article visant à prédire les revenus des films sortant en salle à partir de l'analyse d'un corpus de tweets, Asur et Huberman (2010) écrivent en introduction :

> Since social media can also be construed as a form of collective wisdom, we decided to investigate its power at predicting real-world outcomes. Surprisingly, we discovered that the chatter of a community can indeed be used to make quantitative predictions that outperform those of artificial markets.

Dans l'article, la notion de sagesse collective n'est pas définie, mais on peut déduire de cet extrait qu'elle inclut (ou correspond à) la capacité à faire des prédictions quantitatives. À partir de l'extrait, on peut reconstituer le raisonnement suivant :

- Si les médias sociaux permettent de faire des prédictions quantitatives, alors ils constituent une forme d'intelligence collective.
- Faisons l'hypothèse que les médias sociaux permettent de faire des prédictions quantitatives.
- Pour la tester, on cherche à déterminer si Twitter permet de prédire les revenus d'un film.
- Les revenus prédits par notre expérience sont proches des revenus réels des films
    - donc notre expérience produit un résultat positif.
  - Twitter est un média social et les revenus d'un film sont une variable quantitative
    - donc les médias sociaux permettent de faire des prédictions quantitatives,
    - donc ils constituent une forme d'intelligence collective.

On a donc bien une hypothèse générale, une hypothèse plus spécifique et enfin un cas particulier qui fait l'objet d'une expérience à partir de laquelle on considère que l'hypothèse générale est confirmée. Cet exemple de l'article atteste de l'existence de raisonnement déductiviste dans les *big data*.

En tant que publication scientifique (en l'occurrence pour une conférence internationale en « web intelligence and intelligent agent technology »), l'article renvoie plutôt à ce que Reichenbach appelle le contexte de justification. Ce contexte, sur lequel on reviendra au chapitre VII, qui est celui où le chercheur reconstruit l'expérience scientifique qu'il a réalisée de manière à la présenter comme la preuve de l'hypothèse qu'il formule, s'oppose au contexte de découverte, qui renvoie pour sa part à l'expérience elle-même, c'est-à-dire aux pratiques effectives. Or, de multiples travaux en sociologie des sciences et en communication scientifique ont montré le décalage qu'il pouvait y avoir entre le processus suivi par le chercheur au cours de son expérience et la façon dont il rend compte de ses travaux auprès de la communauté scientifique. Le laboratoire est le lieu du tâtonnement, de l'essai, de l'expérimentation, d'un jeu de circulation entre théorie et expérience, où les hypothèses de recherche ne préexistent pas nécessaire et ne sont du moins pas figées, mais naissent, évoluent, disparaissent, tandis que les expériences sont ajustées, réinterprétées, discutées, etc. Les expériences peuvent être inspirées par les contraintes économiques d'un projet de recherche, l'arrivée d'un nouvel appareil de mesure, l'échange avec un collègue, etc., plutôt que de découler absolument de la formulation d'une hypothèse de recherche précise. Dans ce type de configuration, les raisonnements correspondants peuvent être tout aussi bien inductifs que déductifs. Cette dimension exploratoire des pratiques

scientifiques (par opposition à la dimension confirmatoire qui caractérise davantage le contexte de justification) s'applique également aux pratiques d'analyse de données dans les *big data*.

En l'absence d'une culture épistémique et interprétative adaptée à l'objet analysé, les explications des praticiens des *big data* sont également souvent assimilées à des raisonnements de sens commun, et en particulier à l'abduction telle qu'elle est décrite par Peirce (Rob Kitchin 2014; Vayre 2014; Grolemund et Wickham 2014). L'abduction peircéenne, qui se manifeste à la fois en contexte de découverte et de justification, consiste à adopter (voire inventer) une explication satisfaisante au regard des faits (Douven 2011). La différence avec l'induction est qu'elle ne découle pas logiquement des faits, mais qu'elle est souvent l'explication la plus simple, la première explication acceptable au regard du sens commun. Le raisonnement abductif joue un rôle clé dans les phases d'exploration de données, et se retrouve aussi dans les comptes-rendus d'analyse de données, qu'il s'agisse de rapports produits par des entreprises ou de publications de chercheurs en informatique.

Conclusion

L'inductivisme entretient donc lui aussi un rapport plus complexe avec les pratiques des *big data* que ne le laissent entendre les discours mythologiques associés. Comme la corrélation, l'exhaustivité, l'empiricité, il s'agit davantage de traits projetés et espérés, d'un certain ethos épistémologique, qui n'a rien de nécessaire dans la production de connaissances, et ne s'oppose pas nécessairement à la science traditionnelle même envisagée de manière monolithique.

L'hypothèse selon laquelle toutes ces thèses sont contestables se confirme sans qu'il nous ait été nécessaire de les réfuter entièrement. En tant que telles, les *big data* sont un imaginaire sans définition, une économie de la promesse dont les thèses occupent bien un statut doxastique, avec une fonction rhétorique. La variabilité de leur signification en fait un phénomène culturel qu'il faut aborder en système, sans le décomposer, pour pouvoir l'embrasser d'un seul regard. Néanmoins, au sein de cet imaginaire, nous avons pu dégager des thèses d'ordre épistémologique, sur la nature et le régime de constitution de la connaissance. Nous avons pu voir dans quelle mesure ces thèses sont soutenables de manière générale, c'est-à-dire en quoi ce sont des assertions valides, mais aussi en quoi elles coïncident plus spécifiquement avec les pratiques effectives d'analyse de données massives. En d'autres termes, nous avons évalué leur validité logique et leur correspondance au réel. Il en ressort que les discours sur les *big data* se présentent comme une tentative d'instauration et de légitimation de nouveaux régimes de scientificité. Comme nous allons maintenant le voir, les pratiques sont imprégnées de cette tentative mais ne s'y résument pas. En particulier, nous verrons en effet que la question de la scientificité des connaissances produites n'intervient qu'en second plan de la question de la validité des connaissances et par conséquent des mécanismes de validation et de légitimation des connaissances produites, plus que de l'approche générale. Dans ce but, nous proposons un cadre conceptuel et méthodologique hypothétique, celui des sciences computationnelles de la culture.

Deuxième partie.

Traces et calculs dans les sciences computationnelles de la culture

Chapitre III.
De l'empreinte à la donnée du web : une ontologie de la trace numérique

À travers l'examen des croyances épistémologiques véhiculées par le mythe des *big data*, nous avons vu que celui-ci se construisait presque toujours sur l'opposition entre la science traditionnelle et les nouvelles pratiques d'analyse de données numériques. Pour ce faire, les discours s'appuient sur un sophisme (dit « argument de l'homme de paille ») qui consiste à présenter une vision erronée, volontairement simplificatrice, des habitudes et principes scientifiques, de manière à mettre en valeur la solution proposée par les *big data*. En pratique, on a vu notamment à travers les analyses de Sabina Leonelli et de Carl Lagoze que les sciences de la nature ne sont pas systématiquement menacées ou remises en cause par les promesses des *big data*.

Le fait de manipuler « beaucoup de données » dans le contexte des *data-intensive sciences* représente en premier lieu un enjeu technologique qu'ont bien saisi les équipementiers comme Microsoft. Pour ces équipementiers, la rhétorique des *big data* est utilisée comme discours d'accompagnement du matériel et des logiciels nécessaires au traitement de grands volumes de données scientifiques. Le traitement de données massives soulève également un certain nombre d'enjeux méthodologiques liés à la contextualisation des données et à la nécessaire collaboration entre chercheurs qui se répartissent les compétences nécessaires au traitement, à la manipulation et à l'analyse de ces données (Leonelli 2016).

Ces nouvelles pratiques émergent à côté des pratiques traditionnelles des sciences de la nature et du paradigme galiléen qui les caractérise. En astronomie, en physique des particules, l'informatisation des instruments de mesure génère des masses considérables de données mais dans un cadre théorique globalement inchangé. Dans ce contexte, les *big data* tendent plutôt à renforcer les régimes de scientificité traditionnels en leur apportant de nouveaux observables et de nouveaux outils d'analyse. Il y a en revanche tout un champ de la production de connaissance, qu'elle se dise scientifique ou non, et qui va des pratiques amateurs aux grands projets de génomique, en passant par les logiciels d'analyse de données web et l'analyse littéraire sur ordinateur, qui se voit transformée par les *big*

data. C'est ce nouveau champ que nous allons analyser en lui donnant deux caractérisations : de nouveaux observables, qui ne singularisent pas tant par leur volume que par leur nature même, et de nouveaux outils qui induisent une configuration d'acteurs spécifique.

En termes de typologie d'acteurs en effet, les discours sur les *big data* sont davantage entretenus par les argumentaires des entreprises du secteur informatique, qui commercialisent des solutions de traitement des *big data*, que par des acteurs de la recherche en sciences de la nature. Relayés par les médias, ils ne s'adressent ainsi pas tant aux chercheurs en sciences de la nature qu'à l'écosystème de ces entreprises logicielles : clients, partenaires, candidats. Ils sont particulièrement flatteurs pour les informaticiens (chercheurs, professionnels ou amateurs) amenés à manipuler de grandes quantités de données, qui sont, rappelons-le, présentés par Domingos comme des dieux créateurs d'univers. En ces termes, on comprend que ce type d'acteur soit prompt à relayer et prendre à son compte la rhétorique des *big data*. S'ils ne sont pas forcément le public premier de la mythologie des *big data*, ceux qui doivent y adhérer, ils en sont les héros. Ce statut héroïque fait écho à un état de fait où ils sont la population la plus à même de manipuler des données numériques. Ainsi, leurs compétences techniques en font *de facto* les principaux acteurs des *big data*, acteurs dont les pratiques sont influencées par la rhétorique correspondante : pour ces raisons, il nous était nécessaire de rendre compte des principaux arguments de cette rhétorique, dont le succès permet d'expliquer un certain nombre de croyances et de pratiques.

Néanmoins, nous verrons dans les chapitres qui vont suivre que cette configuration d'acteurs est incomplète. Elle ne suffit pas à produire des connaissances qui émergent dans une double constitution technique et épistémique. Si l'on compare par exemple avec une configuration classique dans les sciences de la nature, où les chercheurs, qui possèdent les concepts théoriques et les connaissances méthodologiques de leurs discipline, s'appuient sur d'autres acteurs qui possèdent les connaissances techniques nécessaires à l'exploitation des instruments de mesure, l'exploitation des *big data* par les seuls informaticiens est une configuration incomplète où les compétences techniques ne sont pas complétées par des connaissances théoriques relatives à l'objet étudié. Il n'y a pas de continuité épistémique ou méthodologique entre la théorie, les modèles et les outils, tout simplement car il n'y a dans cette configuration, pas de cadre théorique. Au regard du fonctionnement classique des sciences, cette configuration n'est pas un « nouveau paradigme » comme le veut la rhétorique des *big data*, mais une situation problématique au sein de laquelle il n'est pas possible de faire émerger de nouvelles connaissances valides.

L'enjeu des chapitres qui vont suivre est donc d'évaluer le rôle de ces compétences techniques, mais aussi de tenter de les inscrire dans une continuité méthodologique qui intègre un cadrage théorique, une conceptualisation de la donnée et des normes de validation des connaissances. En premier lieu, il s'agit de déterminer, en l'absence des normes d'échantillonnage et de représentativité de la statistique inférentielle, comment attribuer une valeur épistémique aux *big data*.

Nous avons vu que la singularité des *big data* par rapport aux pratiques épistémiques antérieures procède essentiellement de la donnée elle-même. En première approche, on peut en effet les considérer comme de nouvelles données qui ne proviennent pas des instruments de mesure scientifiques et sont produites en dehors du cadre des sciences de la nature. Dans les typologies des « *big data* » esquissées dans la littérature universitaire, les données astronomiques ou génomiques par exemple, sont absentes ou anecdotiques. L'analyse de 26 « types » de *big data* proposée par Rob Kitchin et Gavin McArdle (2016) exclut ce type de données. Les sources listées sont les communications mobiles, le web et les réseaux sociaux, les capteurs et caméras, les transactions (comme le fait de scanner un code barre ou d'effectuer un paiement avec une carte bleue), et enfin les administrations. Toutes ces données ont en commun d'être produites par des activités d'origine humaine, et entrent donc difficilement dans le champ des sciences de la nature. Il n'est pas erroné de considérer ainsi que les *big data* sont bien souvent de nouveaux observables pour les sciences de la culture.

Nous allons en effet nous appuyer sur la distinction entre les sciences de la nature et les sciences de la culture, mais il convient dès à présent de la nuancer. En effet, l'étude du vivant, de la santé, les *data-intensive sciences* relèvent des sciences de la nature. La spécificité des *big data* n'est donc pas leur objet mais le statut de leurs observables, dont va découler un tout autre cadre méthodologique que celui des sciences galiléennes. A la mesure des objets du monde, directement analysés dans une continuité théorique qui va de l'instrument à la publication, se substitue une exploitation *a posteriori* de données toujours déjà secondaires, presque toujours déjà matérialisées lorsqu'on envisage de les traiter. Il s'agit donc d'un cadre de travail dans lequel domine l'influence d'une certaine conception des sciences de la culture, mais où n'importe quel objet peut être mobilisé.

A partir de cette conception, que nous allons développer davantage, nous allons examiner dans les chapitres qui suivent les conditions de possibilité des hypothétiques *sciences computationnelles de la culture*, une formule que nous proposons pour désigner un ensemble de pratiques épistémiques combinant un certain cadre théorique et méthodologique développé à partir des sciences de la culture avec la capacité à mobiliser des données numériques massives et des outils de traitement computationnel. Dans la mythologie que nous avons analysée, un élément mérite en effet d'être retenu pour le caractère d'évidence qu'il a pour les praticiens : les *big data* suscitent une complexité technique qui implique des compétences spécifiques et ne permet pas de mobiliser tels quels les outils existants. Les deux composantes essentielles de ces sciences putatives sont 1) la culture épistémique des sciences de la culture (au sens que nous allons leur donner), et leur capacité à conceptualiser un rapport au réel, et 2) la culture technique des informaticiens capables de traduire les concepts et méthodes en outils concrets, compatibles avec le cadre conceptuel préalablement défini. Nous allons montrer d'une part que ces composantes existent, mais d'autre part qu'elles ne parviennent presque jamais à s'articuler, et qu'il manque donc une continuité épistémique entre ces deux composantes.

D'un côté, la problématisation du statut de la donnée et de sa valeur épistémique est émergente dans plusieurs communautés de recherche. Techniquement mis en difficulté par les nouveaux modes d'accès au réel qui pourraient s'offrir à eux, car ils n'ont pas les compétences dont disposent les informaticiens, ces chercheurs sont en revanche sensibles aux problèmes épistémologiques qu'ils posent. Si ce n'est résolus à les résoudre, ils sont du moins théoriquement convaincus qu'un problème existe. Si ces acteurs parvenaient à se doter de moyens techniques concrets pour traiter des données, cette démarche pourrait se faire au sein d'une culture épistémique homogène articulant les problèmes de recherche, les données, les outils et les méthodes.

De l'autre côté, les mêmes informaticiens possèdent la capacité concrète à manier de grands volumes de données hétérogènes, ou à développer les objets requis, mais ne s'inscrivent pas, par définition dans l'épistémologie des sciences de la culture. Leur intervention prend la forme de manipulations régies non pas par un projet épistémique relatif au fait humain, mais par l'exploration systématique de l'espace de manipulabilité apporté par l'informatique. Nous reviendrons et confirmerons en détails cet état de fait au chapitre V.

Avant cela, nous allons expliciter quel statut on peut donner aux observables dans les sciences de la culture, et la façon dont ces sciences construisent un rapport au réel. Dans cette perspective, nous verrons ainsi quelle conceptualisation de la donnée numérique peut être proposée dans le champ des sciences de la culture, et quelle redéfinition des sciences de la culture elles-mêmes est induite par l'irruption de ces nouveaux observables.

1. L'épistémologie des sciences de la culture

Avant de développer quel rapport au réel et quelles normes régissent les sciences computationnelles de la culture, il nous faut préciser l'origine et le sens de ce terme, par rapport à d'autres découpages disciplinaires notamment. La notion de « sciences de la culture » nous vient ainsi de l'école néokantienne de Heidelberg, incarnée notamment par Rickert et Windelband ; ce dernier le précède historiquement et parle plutôt de « sciences de l'esprit », mais dans un sens globalement identique. Chez Windelband (2000), le projet de sciences de l'esprit induit l'idée qu'il y a plusieurs façons d'aborder philosophiquement la question de la culture : une approche normative, à travers une philosophie de la culture qui s'efforcerait de fonder une norme universellement valide pour une culture à venir, et une approche descriptive, à travers une philosophie des sciences de la culture, qui réfléchit à la manière dont on peut étudier empiriquement les cultures effectives et fonder en science cette étude empirique. Pour un néokantien comme Windelband, cette distinction entre normatif et descriptif renvoie à la séparation kantienne entre l'éthique de la *Critique de la raison pratique* et l'épistémologie de la *Critique de la raison pure*. Chez Kant, il n'y a de science que de la nature, tandis que la culture s'envisage sous l'angle des valeurs et de la subjectivité pure : la culture est non pas un

état de l'humanité mais le processus de se cultiver, un processus considéré comme un devoir de l'homme, qui lui est anthropologiquement constitutif (se cultiver est ce qui nous rend humains).

L'un des projets des néokantiens est précisément de prolonger (ou oserait-on dire, améliorer) l'œuvre kantienne en faisant émerger et en légitimant le projet des sciences de la culture, c'est-à-dire d'une connaissance objective, non normative, de la culture. La philosophie de la culture ne serait plus seulement une philosophie interne du sujet, mais aussi une philosophie transcendantale de la culture en tant qu'objet. Cette philosophie, dans l'ébauche qu'en font les néokantiens, montre que les sciences de la culture se caractérisent à la fois par leur méthode et par leur objet. Les sciences de la nature désignent à la fois leur méthode (la naturalisation des phénomènes et la recherche de lois) et leur objet (les phénomènes naturels) : pour être absolument rigoureux, il faudrait parler de sciences naturelles de la nature. Rickert (1997) préfère parler de sciences *historiques* de la culture, c'est-à-dire de sciences dont la méthode est historique et l'objet la culture, et souligne qu' « il nous manque un terme, équivalent de celui de "nature", qui les désignerait tout autant par rapport à leur objet que par rapport à leur méthode ». Ce couplage entre méthode et objet n'est pas systématique puisque la distinction même entre sciences de la nature et sciences de la culture est plutôt un spectre qui sert à « présenter les deux *extrêmes* entre lesquels se situent *presque tous les travaux scientifiques* » (*Ibid.*).

Il existe ainsi au moins une science naturelle de l'esprit (la psychologie, chez Windelband). Par ailleurs, les sciences de la vie occupent un statut spécifique vis-à-vis de ce spectre : l'étude du vivant implique de se donner le concept de fin pour comprendre la fonction d'un organe, d'un trait génétique, mais ce concept est seulement régulateur dans la mesure où il n'y a pas de finalité dans le vivant. Tout se passe comme si l'étude du vivant s'inscrivait dans un régime compréhensif envisagé comme fiction méthodologique ; il est possible, par ailleurs, de naturaliser l'étude du vivant en quittant l'échelle de l'organisme pour aller vers des objets plus spécifiques tels que la cellule ou la matière organique. L'exemple du vivant montre que la dichotomie entre sciences de la nature et sciences de la culture n'est pas mutuellement exclusive. Néanmoins, cette distinction constitue une clé d'intelligibilité pour situer épistémologiquement des travaux relevant de ce que nous appelons aujourd'hui les sciences humaines et sociales, et tout particulièrement de l'histoire.

Les sciences historiques de la culture ont pour objet les événements culturels, c'est-à-dire les résultats des activités de l'homme en tant qu'il vise une fin. Elles s'intéressent à l'événement dans sa particularité et dans son individualité, à ce qui le rend unique et singulier dans l'histoire ; elles sont dites individualisantes ou *idiographiques*, c'est-à-dire qu'elles reposent sur l'écriture du singulier. Rien n'interdit d'imaginer des sciences historiques de la nature qui mettraient en évidence la singularité d'une feuille d'arbre, d'un bloc de granit, d'un amas de cellule ; il existe par exemple des approches historiques en biologie retraçant l'évolution historique d'un organisme d'une génération d'individus à une autre. La pratique médicale, fondée sur des connaissances biologiques générales, examine un patient dans sa singularité, avant de ramener les symptômes identifiés à des connaissances plus

générales. Néanmoins la plupart des sciences de la nature s'en tiennent à la méthode naturelle, aussi dite généralisante ou *nomologique*, qui s'efforce d'identifier des lois, ou du moins des relations générales entre les phénomènes. Parmi nos disciplines actuelles, les sciences historiques de la culture correspondraient plus naturellement à l'histoire générale (sauf l'histoire sérielle), l'histoire de l'art, l'histoire des religions, les études littéraires, et peut-être aussi l'anthropologie en tant qu'elle vise un groupe humain dans sa singularité.

À l'inverse, les sciences sociales quantitatives telles que la sociologie ou l'économie, inexistantes dans leur forme actuelle à l'époque de Windelband et de Rickert, font plutôt le pari d'une méthode nomothétique appliquée à l'étude d'objets culturels et sociaux, avec une épistémologie de la mesure où l'observation et la collecte de données scientifiques permettent l'émergence de régularités statistiques et de lois. Néanmoins, comme on le verra ci-dessous, elles présentent un ensemble de spécificités lié à la singularité de leur objet, et notamment au fait que la naturalisation des faits humains n'épuise pas la capacité de l'homme à s'envisager comme fin (Schurmans 2011). Cette capacité place la recherche d'explication et de compréhension en sciences sociales quantitatives dans le domaine des raisons (plutôt que des causes) ou condamne le chercheur à rester à la surface phénoménologique du fait humain. Le chercheur se donne lui-même pour objet, rendant artificielle et purement méthodologique la distinction entre sujet et objet : qu'il ressemble ou diffère des groupes humains qui constituent son objet, il est la mesure de toute chose connaissable dans l'espace du fait humain, envisagée à l'aune de sa subjectivité irréductible. De là, deux grandes directions sont possibles pour lui : une recherche de scientificité à travers un objectivisme phénoménologique qui neutralise, ou mieux, intègre, la réflexivité que son objet présente, ou le maintien de la subjectivité dans un pluralisme des interprétations du fait humain, évalué sous d'autres critères que celui de la scientificité propre aux sciences nomothétiques de la nature.

2. La trace dans les sciences indiciaires

Nous allons donc envisager les *sciences computationnelles de la culture* comme une certaine forme de sciences de la culture qui, conformément à la conception de Rickert, n'a pas un objet spécifique, mais se caractérise en revanche par sa méthode : les sciences computationnelles de la culture sont définies comme telles en raison de leur mode d'accès au réel, et d'une certaine façon d'aborder les objets. Ce mode d'accès prend une forme classique dans les sciences historiques de la culture ; néanmoins, il est également en usage dans d'autres contextes comme celui du diagnostic médical ou de l'enquête policière. Ce mode d'accès s'agit de celui de la trace ou indice, qui mobilise un paradigme épistémologique indiciaire. La trace sert de source à l'historien pour lequel elle atteste des faits passés qui ne sont, par définition, plus accessibles. Les témoignages, les vestiges et les objets qui ont perduré à travers le temps, documentent ce qui n'est plus. Chez Paul Ricoeur, la trace (ou plus précisément, la trace documentaire, par opposition à la trace affective ou corporelle) est à la connaissance historique

ce que l'observation directe ou instrumentale est aux sciences de la nature (Serres 2002), autrement dit son fondement empirique. Plusieurs traits fondamentaux distinguent la trace de la mesure produite par les instruments scientifiques : son existence n'est pas la conséquence des actions de celui qui cherche à l'étudier, de sorte qu'elle possède toujours une originarité qui échappe à son observateur. Elle s'inscrit également toujours dans une dimension langagière (qu'il s'agisse d'un langage symbolique sommaire ou d'une écriture) à laquelle les concepts et méthodes des sciences de la culture, et notamment des sciences du langage, sont naturellement reliés.

Dans le contexte du web et des réseaux sociaux, la trace numérique est une notion historiquement attestée, déjà chargée de sens. Elle est en quelque sorte le pendant des *big data* pour les chercheurs en sciences de la culture : une notion médiatique, floue, axiologiquement chargée, davantage problématisée d'un point de vue éthique et politique car « le plus souvent ramenée à une opposition entre protection et exhibition de la vie privée » (Merzeau 2013a) mais également rattachée à des traditions épistémologiques, dont tout naturellement celle de la trace tout court, qui mérite donc d'être conceptualisée ici.

On se perd aisément à énumérer les nuances de sens du mot « trace », qui peut être indice, empreinte, petite quantité. La comparaison entre la trace de pas laissée dans la neige et l'archive de l'historien facilite la compréhension mais non la définition. Conceptuellement, nous dirons que la trace est le résidu matériel d'un événement toujours déjà passé. Elle est ce qui demeure de l'irruption d'une causalité hétérogène dans un environnement régi par ses propres causes : ainsi le pas dans la neige résulte d'une part du système physique de la neige, de la façon dont elle tombe, s'agrège, se déforme, se maintient, et en l'occurrence, est imprimée par un corps, et d'autre part, de l'irruption d'un autre système causal, celui du corps du promeneur ou de l'animal déclencheur de l'événement, du pas dans la neige. La trace a toujours une matérialité qui lui permet de persister et de jouer le rôle de preuve. Bien que cette signification soit, comme on le verra, constitutive de la trace, il y a une positivité irréductible de la trace qui ne se limite pas à la signification qui peut lui être conférée : elle est toujours à la fois forme matérielle et signe.

Cependant, sa mobilisation épistémique ne s'intéresse pas à cette matérialité en tant que telle ; cela reviendrait sinon à étudier l'index proverbial quand celui-ci désigne la lune. Comme le souligne justement Alexandre Serres (2002) :

> Comme d'autres termes généraux et complexes (comme celui de forme notamment), la trace se caractérise par son génitif intrinsèque, si l'on peut dire, i.e. son caractère d'appartenance, au sens où la trace est toujours trace de quelque chose ; elle ne se définit pas par elle-même, elle n'a pas d'existence propre, autonome, au plan ontologique du moins, elle n'existe que par rapport à autre chose (un événement, un être, un phénomène quelconque), elle est de l'ordre du double, voire de la représentation et ne prend son sens que sous le regard qui la déchiffrera. D'où une certaine difficulté, sinon à définir du moins à caractériser et surtout à inventorier les traces, puisque tout peut devenir trace de quelque chose.

Leur utilisation dans un contexte épistémique se comprend donc mieux à partir de leur fonction que de leur nature et de leur matérialité, qui est la condition nécessaire mais pas suffisante de leur « tracéité ». Les traces sont des objets sémiotiques qui peuvent prendre n'importe quelle forme pourvu qu'elles représentent quelque chose. Dans la perspective de l'étude des traces numériques, on ne cherche donc pas tant à comprendre la forme matérielle que peut prendre la trace qu'à examiner comment elle fonctionne et ce qui fait qu'elle peut être exploitée. En ces termes, se dégagent plusieurs conditions de possibilité d'une utilisation épistémique des traces à partir de leur fonction représentationnelle, et dont on verra qu'elles ne sont pas à prendre telles quelles pour ce qui est des traces numériques.

D'une part, la trace doit provenir d'un couplage attesté avec l'événement qu'elle désigne pour jouer son rôle de preuve. Du fait qu'une trace peut souvent être falsifiée, elle n'est acceptable comme preuve qu'à partir du moment où sa propre traçabilité peut être attestée, ou du moins raisonnablement supposée, par exemple avec une probabilité ou une confiance élevée : les discours d'accompagnement des mobilisations épistémiques de traces servent alors à justifier la confiance qui est placé dans la trace en question. Elle ne peut jouer le rôle de preuve de quelque chose que si elle apporte d'abord la preuve de sa propre authenticité. Dans la tradition archivistique, le document est envisagé comme la trace matérielle de l'événement, avec lequel elle entretient un rapport organique, quasi-mécanique. Une manière efficace de rendre ce couplage effectif est de le produire volontairement. À ce titre, l'acte de vente notarié constitue un exemple emblématique de la trace. C'est un document que le notaire s'efforce de rendre infalsifiable parce qu'il doit être quasi-performatif de la vente effective : dans l'idéal notarial, il y a l'événement « vente » non pas simultanément à l'acte notarié mais *en conséquence de* l'acte notarié qui en est la condition nécessaire et suffisante. Faire trace est ici un acte intentionnel, dont la trace est la conséquence, et où le couplage entre la trace et l'événement résulte d'un effort particulier, d'une intentionnalité rendue effective par un ensemble d'actions.

Néanmoins, cette intentionnalité de la trace n'est pas systématique. Ce qui caractérise systématiquement la trace, en revanche, c'est d'être le résidu matériel irréductible d'un événement passé. De fait, du point de vue non pas de l'individu qui produit la trace, mais de celui qui l'examine, la signification de la trace échappe au moins en partie à son auteur, un peu à la façon dont le livre échappe à son auteur et construit sa signification par ses lecteurs : il y a bien une intentionnalité d'action initiale, mais à partir de laquelle un déplacement d'intentionnalité se produit dès lors que cette trace est interprétée par un lecteur. Ainsi, chez Sybille Krämer, la trace doit être involontaire, laissée à l'insu de la personne ou de la chose qui la produit :

> On ne fabrique pas une trace, on la laisse, et ce sans intention aucune. De même, effacer des traces revient à en laisser une. Et vice versa : dès lors qu'une trace est sciemment laissée et mise en scène en tant que telle, il ne s'agit plus d'une trace. Seul ce qui n'est pas intentionnel, ce qui est involontaire, incontrôlé, arbitraire,

> grave ou dessine ces lignes de rupture qui peuvent être lues comme des pistes. À la différence du signe que nous créons, la signification d'une trace existe au-delà de l'intention de celui qui la génère. (Krämer 2012)

Qu'elle soit involontaire ne veut pas dire qu'elle ne peut pas être d'origine intentionnelle. Ainsi, si je dessine une marque dans le sol dans l'intention de laisser une trace, la marque n'est pas tant la trace d'elle-même que la trace de mon intention de laisser une trace. Du point de vue de celui qui l'observe, la trace s'interprète toujours à un niveau d'intentionnalité autre que celui de l'auteur de la trace : l'épistémologie de la trace est une épistémologie du lecteur de traces.

Enfin, une trace n'est telle que si elle a une signification pour un observateur qui est en mesure de l'interpréter. Elle ne doit pas seulement être perçue (directement ou indirectement, à travers un instrument ou un médium), mais comprise. N'est pas trace tout ce qui est perceptible, mais seulement ce qui est identifié comme tel par une « perception sélective de l'environnement » (*Ibid.*). Si toutes les traces ont une positivité matérielle par laquelle elles ne sont pas purement construites par sélection, une trace n'en est pas moins le résultat d'un processus de sélection qui singularise (ici au sens de « détacher » plus que de « rendre unique ») un élément par rapport à son environnement. Pour le pisteur, la branche cassée qui signale le passage du cerf est singularisée non seulement par rapport à toutes les branches de la forêt, mais aussi par les branches cassées accidentellement, par dessèchement, par le vent, ou toute autre cause qui n'est pas le passage du cerf. Celui-ci n'est pas singularisé en tant que tel, mais signalé par la trace elle-même singularisée, différenciée de son environnement. Cette singularisation ne préexiste pas au pisteur, elle résulte de son talent d'observation et d'interprétation.

Or, qui dit interprétation, dit mobilisation d'une culture herméneutique, d'un savoir-faire relié à d'autres connaissances mobilisées en contexte, ou autrement dit d'un « habitus permettant de viser la trace comme trace, une tradition dans laquelle on maintient l'interprétabilité de la trace » (Bachimont 2010a). La trace est donc toujours trace *de* quelque chose mais aussi trace *pour* quelqu'un ; tout peut être trace car la « tracéité » de la trace procède d'un couplage entre un reliquat matériel et sa structure interprétative, entre sa positivité matérielle et sa singularisation intentionnelle constitutive.

La trace serait-elle donc la donnée des sciences humaines ? L'interprétation dont elle dépend n'est pas une science mais un art auquel on connaît de multiples pratiques, de l'haruspice au médecin en passant par la compréhension intuitive dont nous faisons usage au quotidien. Que les sciences de la nature présentent une dimension herméneutique ne fait pas de doute, quoique l'interprétation du physicien des particules puisse être plus formellement codifiée que celle de l'historien. Ce n'est donc pas le caractère herméneutique de la trace qui la distingue de la donnée des sciences de la nature. Lorsque Carlo Ginzburg (1980) oppose la science galiléenne au paradigme indiciaire du médecin, du détective ou du critique d'art, il souligne la différence entre leurs modes d'appréhension de l'objet :

d'une part de l'universel répétable, de l'autre de l'individuel singulier. La donnée des sciences de la nature repose sur la répétabilité des phénomènes, leur quantité, les critères sous lesquels ils peuvent être considérés comme semblables. Les disciplines herméneutiques s'intéressent au contraire à la singularité du phénomène, à ce qui caractérise *ce* texte, *cette* personne, *cette* image – vers laquelle je fais signe en la désignant. Opter pour une épistémologie indiciaire de la trace, c'est, dans les termes du néokantisme de Heidelberg, s'inscrire dans une perspective idiographique plutôt que nomologique (Rickert 1997; Windelband 2000), cherchant à comprendre la signification culturelle de la trace dans sa singularité, plutôt qu'à expliquer les mécanismes causaux dont procède son existence.

Nous proposons donc de considérer les données comme le pluriel de la trace. Ce passage au pluriel n'est pas une pirouette grammaticale ; il se fait au prix d'un renoncement à cerner qualitativement ce qui fait la singularité de la trace et à condition de trouver ce qui permet d'envisager les traces comme un collectif susceptible de présenter des régularités. Subsumer la trace sous le général, c'est faire le deuil de la richesse inépuisable de *l'*événement singulier pour enregistrer le caractère commensurable *des* événements et leur reproductibilité dans le temps et l'espace. De cette reproduction émerge le général, non pas comme un trait systématique, mais quantitativement majoritaire. A partir du général, des singuliers, des écarts à la norme, pourront être reconstitués. Enfin, quel que soit leur régime d'appréhension, la donnée envisagée comme le pluriel de la trace est toujours initialement privée de son originarité et de la connaissance de ses conditions d'émergence.

3. Le log ou journal d'utilisation

Si l'on revient maintenant en environnement numérique, on constate bien que les traces numériques se présentent toujours déjà comme un pluriel (on ne dit pas « une » mais « des » traces numériques). Le numérique étant le terrain par excellence de la répétition et de la commensurabilité formelle, pourquoi serait-il encore nécessaire de favoriser le terme de trace, lié à la notion de singularité ?

La raison première est historique et pratique. Les traces numériques renvoient d'abord, en première acception, à un dispositif informatique permettant à un développeur de diagnostiquer son propre travail. En écrivant son programme, il définit également comment les différents processus produisent, en s'exécutant, des traces – les *logs*, qui n'ont pas de bonne traduction en français – grâce auxquelles il va pouvoir *débugger* son programme, c'est-à-dire identifier la source des *bugs* ou erreurs qu'il produit. Comme l'écrivent Champin, Mille et Prié (2013) :

> Les traces numériques ont été utilisées d'emblée pour faciliter le débogage de programmes informatiques avec l'idée qu'un observateur averti (le programmeur en général), analyste de l'observation faite, sera capable d'interpréter les traces issues de l'exécution du programme pour en comprendre le comportement et le corriger en cas de besoin.

Les traces informatiques[13] n'ont donc pas de visée de connaissance en tant que telle, mais de contrôle (au sens de vérification et non de pouvoir sur) ; il ne s'agit pas de mieux comprendre l'utilisateur mais de vérifier que le programme fonctionne de la façon prévue. Contrairement à la mobilisation de l'archive par l'historien, elles ne relèvent pas tant d'un projet scientifique que d'une problématique d'ingénierie. Elles sont toutefois, comme la trace « analogique », un mode d'accès à la singularité de ce qui les a produites, et devront être interprétées pour reconstituer et comprendre l'événement dont elles sont le témoignage. Elles sont bien des irruptions d'une causalité hétérogène dans un environnement régi par ses propres normes causales. L'environnement est le système d'enregistrement des événements régi par le programme, et la causalité est celle qui est capturée par ce système sous forme d'une entrée dans les logs : une interaction utilisateur, un dysfonctionnement, la complétion d'une tâche, etc. La différence essentielle est celle du médium : le coup de pinceau, la fièvre, l'empreinte digitale, sont des vestiges matériels de l'impression d'un objet (ou agent) sur un autre, tandis que la trace numérique est une empreinte informationnelle générée par un artefact computationnel. De ce fait, contrairement à l'acte notarié où la production de la trace est le résultat d'un processus intentionnel, la trace numérique procède d'un couplage mécanique entre le système d'inscription et l'événement informatique. Autrement dit, dans un cas la trace en tant que telle est interprétée comme le résultat d'une intention ; dans l'autre le couplage entre l'événement informatique et sa cause « analogique » (c'est-à-dire, pour simplifier, l'utilisateur du programme) est reconstruite par le lecteur lorsqu'elle existe.

Par ailleurs, ce passage d'un substrat matériel à un substrat informationnel a quelques conséquences :

- toute trace numérique peut être horodatée (écriture d'un *timestamp)* ;
- la trace doit être discrétisée par un jeu de symboles (caractères, pixels, etc.) et suivant un encodage spécifique ;
- elle devient toujours déjà manipulable et calculable par nature (*Ibid.*).

Si les logs sont primairement des outils de diagnostic du comportement du programme lui-même, ils deviennent aisément des outils de contrôle du comportement de l'utilisateur devant le programme. Les traces du programme sont alors réinterprétées comme des traces de l'utilisateur du programme. Dans cette configuration, il y a une mise à distance de la situation d'utilisation par rapport à la situation de contrôle. La machine sur laquelle le programme est exécuté par l'utilisateur n'est plus forcément celle où le comportement de l'utilisateur est étudié[14]. Concrètement, trois configurations peuvent se présenter selon le mode d'installation du logiciel :

[13] On notera le passage de « trace numérique » à « trace informatique ». Le constat est que la formulation varie en fonction du locuteur : en informatique, on parle généralement de trace informatique, et en SHS de trace numérique, les chercheurs en SHS s'intéressant davantage à la signification culturelle de l'objet qu'à ses modalités techniques. Or l'adjectif « numérique » renvoie généralement à la dimension culturelle de l'informatique.

[14] En termes de division du travail au sein des organisations, les fonctions de conception et de contrôle sont aussi souvent séparées : le développeur qui écrit le programme n'est pas l'analyste qui étudie le comportement des utilisateurs ; ce

- l'utilisateur installe un logiciel sur sa propre machine. Le logiciel fonctionne en autonomie sur cette machine. Il produit ou non des logs qui restent sur la machine de l'utilisateur ;
- l'utilisateur installe le logiciel sur sa propre machine mais le logiciel envoie des informations à l'organisation qui le fournit. Les logs sont produits sur la machine de l'utilisateur mais analysés dans l'organisation ;
- l'utilisateur n'installe pas un logiciel, mais utilise une application directement en ligne, typiquement dans un navigateur web. Des logs peuvent être enregistrés sur sa machine, ou directement sur le serveur qui héberge l'application web.

Ces différentes configurations matérielles induisent des options épistémiques qui peuvent être différentes : lorsque l'accès aux logs reste localisé sur la machine de l'utilisateur, les consulter sert à examiner l'utilisation ou le fonctionnement du logiciel dans sa singularité, par exemple pour identifier un problème dans l'installation du logiciel lié à la configuration de la machine. Lorsqu'au contraire, les logs sont examinés sur une autre machine, typiquement les serveurs mis à disposition par les éditeurs du logiciel, celle-ci agrège généralement les logs de tous les utilisateurs du logiciel de par le monde. De ce fait, ce n'est plus la singularité d'une utilisation qui est examinée, mais les régularités et irrégularités de toutes les utilisations du logiciel. On passe ainsi du particulier au général, de l'unique au statistique, avec des traces numériques envisagées sous le mode de la comparaison, de manière à dégager d'une part des mesures globales qui vont se constituer en normes, et des mesures plus locales qui vont caractériser des différences à la norme. L'apparition de normes, de comparaisons, de calculs, marque le passage du log envisagé comme trace idiographique du comportement de l'utilisateur ou du programme, aux logs considérés comme données, c'est-à-dire représentations nomologiques, ou du moins créatrices de quasi-lois, de ces comportements.

Ce passage aux logs envisagés dans leur pluralité se double également d'une distinction sociologique ; il faut se demander qui parle, quel agent épistémique intervient sur les logs pour déterminer à quoi l'on a affaire. Du fait de la division du travail entre les fonctions d'ingénierie et de marketing notamment, c'est rarement la même personne qui développe un logiciel et en analyse les usages ; lorsque cela arrive, elle endosse des rôles épistémiques distincts et difficilement simultanés, selon la tâche sur laquelle elle travaille à un moment donné. L'informaticien fabrique des logs pour l'assister dans la vérification de son programme, tandis que l'analyste manipule des données pour mieux comprendre le comportement de ses utilisateurs. Les uns sont en fin de compte des modalités de constitution d'un artefact computationnel, les autres un mode d'accès au réel *via* une représentation statistique. On retrouve ainsi la perte d'originarité constitutive de la trace, car le développeur qui a créé le système générant des traces n'est pas celui qui les analyse et les interprète.

dernier est souvent, en contexte industriel, rattaché à des fonctions marketing. Ils peuvent se trouver géographiquement séparés, voire appartenir à des organisations différentes et ne jamais échanger.

La distinction entre contrôle de l'exécution du programme et analyse du comportement des utilisateurs est cependant loin d'être binaire. L'exploitation de logs renvoie à une multiplicité de cas d'usage avec des projets épistémiques variés, exécutés avec plus ou moins de succès. La mise en place de logs résulte d'une multitude de décisions et repose sur un savoir-faire issu de la pratique et du partage d'expérience. Les développeurs débutants vont avoir tendance à prévoir peu de logs, qui produiront des informations parfois sibyllines, peu contextualisées, qui ne faciliteront pas ou peu la localisation du problème. Un développeur plus expérimenté saura se projeter à l'avance dans la situation de diagnostic pour anticiper les informations de contextualisation à faire figurer dans les logs : quelle action, quel événement, quel processus, quels paramètres, etc. En phase de fonctionnement (par opposition à la phase de conception où le développeur peut procéder à des tests sur sa propre machine), les logs peuvent servir à surveiller une multitude d'objets au sein d'un système d'information : les programmes et les actions des utilisateurs, mais aussi les serveurs, les transferts d'information, le système d'exploitation, les programmes, les tâches planifiées, les bases de données, etc. Ainsi, la distinction idiographique/nomologique ne se superpose pas à l'opposition programme/utilisateur, l'analyse du comportement de l'un ou de l'autre pouvant se faire sous les deux termes de la distinction. De ce fait, on ne peut pas réduire l'analyse des logs à l'analyse des comportements des utilisateurs, qui est généralement la fonction attribuée à l'étude des traces numériques : tous les logs ne sont pas des traces numériques et, comme on le verra, toutes les traces numériques ne sont pas des logs.

De plus, l'analyse des comportements utilisateurs par les éditeurs de logiciel produit des connaissances sur les usages d'un logiciel en particulier, qui sont difficilement généralisables aux relations entre informatique et utilisateurs dans l'absolu, et ne sont, par ailleurs, pas une fin en soi. Cet examen est un projet épistémique au service de la visée marketing et commerciale de l'entreprise, qui ne cherche à comprendre ses utilisateurs que pour les conserver, ou en attirer davantage ; si d'autres motivations peuvent exister, comme par exemple la curiosité intellectuelle, c'est au niveau individuel qu'elles se présentent.

4. La trace numérique comme log du web

Le type d'analyse de logs présenté jusqu'à maintenant est un cas particulier, mais néanmoins originaire, de l'analyse des traces numériques. Du point de vue des connaissances qu'elles peuvent produire, la question des traces numériques a aujourd'hui une portée qui dépasse largement l'examen des utilisations d'un logiciel par ses éditeurs. Les traces numériques sont abondantes en qualité comme en quantité : il en existe de toutes sortes et elles sont produites massivement. L'informatisation d'un grand nombre de secteurs d'activité économique (le commerce, la banque, les transports, etc.), le développement de l'informatique de masse et l'apparition du web, entre autres, ont fait de l'enregistrement des traces numériques un phénomène suffisamment massif pour susciter l'intérêt

d'autres organisations que les entreprises qui éditent des logiciels. L'analyse des flux financiers, des comportements d'achat, des conversations instantanées, pour ne citer que quelques exemples d'activités qui produisent des traces sur des supports informatiques, a une portée qui dépasse largement l'étude par une entreprise de ses utilisateurs. De fait, c'est tout un univers de possibles qui s'ouvre à partir du moment où l'on confère aux traces numériques une certaine valeur de généralité, voire d'universalité. En particulier :

- du fait que l'informatique est le support même de certaines de ces activités, leur enregistrement en produit une représentation intégrale (selon un certain point de vue du moins). Par exemple, l'ensemble des transactions financières des banques est aujourd'hui informatisé ; par conséquent, l'analyse des traces de ces transactions revient à analyser la totalité de son activité en matière de transactions ;
- par leur montée en généralité, les traces numériques deviennent un mode d'accès à l'étude de phénomènes culturels. En France par exemple, la plupart des individus étant équipés d'un téléphone portable[15], l'analyse de leurs traces d'utilisation mobile devient une source pour l'étude des déplacements, des habitudes de communication, de la structure du réseau social d'une personne, etc. ;
- enfin, par leur caractère systématique, elles deviennent pour leurs auteurs un enjeu d'expression, de réappropriation, voire de construction de soi.

Le cas le plus notable de cette abondance nouvelle des traces est le web. Dans la littérature sur les traces numériques, c'est presque toujours de lui qu'il s'agit, même quand cela n'est pas précisé. Le caractère d'évidence du web, phénomène majeur et connu de tous, en fait l'exemple par excellence de dispositif technique producteur de traces numériques. L'usage courant de la notion de trace numérique a ainsi tendance à exclure les traces des utilisations des téléphones portables, des capteurs et objets connectés, des programmes connectés à Internet sans passer par le web, et de tous les réseaux privés tels les dispositifs internes des organisations. Par opposition à d'autres dispositifs du même genre, il présente des qualités qui expliquent au moins en partie ce caractère d'évidence :

- il s'appuie sur un réseau mondial utilisé par près d'un individu sur deux sur Terre, ce qui en fait le plus grand réseau de communication qui ait jamais existé ;
- il bénéficie de ce fait d'une notoriété difficilement égalable ;
- contrairement aux systèmes d'information des banques, il présente, pour d'autres acteurs que les banques elles-mêmes, un certain niveau d'ouverture qui permet, avec plus ou moins de complications techniques, d'accéder aux traces laissées par ses utilisateurs ;

[15] 92% de la population en 2015, d'après le CREDOC, enquêtes «Conditions de vie et Aspirations » (vague de juin de chaque année), http://www.arcep.fr/uploads/tx_gspublication/CREDOC-Rapport-enquete-diffusion-TIC-France_CGE-ARCEP_nov2015.pdf (p. 22, consulté le 16 novembre 2017)

– il est le support d'usages et de comportements culturels suffisamment riches pour qu'un répertoire exhaustif soit une entreprise impossible.

Par l'intensité de l'usage du web, il est techniquement possible d'étudier un grand nombre de traces : leur matérialité numérique leur confère toujours déjà une manipulabilité technique et ouvre un domaine de possibles au sein duquel des manipulations épistémologiquement souhaitables peuvent se dégager. Au regard des traces du web, plusieurs typologies ont été proposées, selon qu'elles sont privées ou publiques, indexables par des moteurs de recherche, textuelles ou relationnelles, primitives ou calculées, etc. Qu'il s'agisse des genres de traces ou des typologies elles-mêmes, il est difficile de parler de traces d'usage du web sans énumérer des listes « à la Prévert » qui permettent seulement de rendre compte de leur diversité sans pouvoir mesurer leur étendue, mais qui ont néanmoins l'avantage de ne pas laisser croire à une forme d'ordre ou d'homogénéité du phénomène. La tendance inverse est en effet à la réduction de l'extension du domaine de la trace, notamment au prisme d'une notion spécifique qui présente un intérêt méthodologique. Il est difficile de discerner entre une simplification méthodologique, c'est-à-dire une modélisation du phénomène au service d'une hypothèse de recherche et qui permet de le rendre intelligible, et une réduction ontologique qui modifie la définition même de ce qu'est une trace en fonction de l'hypothèse de recherche et aboutit à une circularité du raisonnement sur l'objet. Ces typologies sont confrontées au problème de toute classification du monde, nécessairement réductrice, et simultanément constitutive de notre manière de le comprendre. On trouve dans la littérature sur les traces numériques de telles simplifications dont le statut épistémologique n'est pas explicité, et dont on ignore donc si ce qui est décrit désigne la totalité de la notion de trace ou seulement un angle particulier, ouvert à un pluralisme d'approches variées du sujet. Parmi les travaux concernés, citons-en quelques-uns :

– Antonio Casilli (2012) propose d'aborder les traces numériques au prisme de leur dimensionnalité, qui peut être unique (contenu textuel court), double (image, vidéo, texte mis en forme) ou triple (immersion visuelle en 3D) ;
– Dominique Cardon (2013) résume les traces d'utilisation du web à celles des réseaux sociaux, notamment Facebook, et des moteurs de recherche, essentiellement Google ;
– Fanny George (2009) utilise pour sa part la notion d'identité, les traces pouvant être des représentations d'une identité déclarative, agissante ou calculée ;
– Louise Merzeau (2013b) classe les traces numériques en déclaratives, comportementales, ou documentaires en fonction d'une échelle d'intentionnalité de la constitution de la trace.

Bien que la classification de Louise Merzeau nous semble réductrice, nous partageons l'intérêt qu'elle porte à la question de l'intentionnalité, qui constitue l'un des nœuds conceptuels de la notion de trace en général et de trace numérique en particulier.

5. L'intentionnalité des traces numériques

D'après la conception de Sybille Krämer évoquée plus haut, une trace ne peut être que non intentionnelle, il en va de l'essence de la trace que d'avoir été produite involontairement et d'être soumise à l'interprétation du lecteur de trace. Or, Louise Merzeau souligne avec justesse que de nombreux contenus en ligne conceptualisés comme des traces sont au contraire tout à fait intentionnels, et participent même à une activité réflexive de construction du soi. En conséquence, l'essence de la trace numérique n'est pas strictement intentionnelle ou non intentionnelle. L'intentionnalité de nos traces numériques existe par degré, sur le mode du manifeste de soi comme sur le mode du lapsus ; la conscience d'être tracé existe elle aussi par degré. La configuration d'un profil, ou la discussion sur les réseaux sociaux, sont des actes de communication qui permettent de se décrire de façon autonome ou en réaction au regard d'autrui : ce sont des actes parfaitement intentionnels. Dans le même temps, l'enregistrement de la trace est toujours déjà actif, qu'on le veuille ou non :

> Entièrement automatisée, cette traçabilité n'est pas une couche documentaire qui se greffe après coup. Elle est la condition même de la performativité numérique. De la même façon qu'on ne peut pas ne pas communiquer, on ne peut pas ne pas laisser de traces. (Merzeau 2013a)

Aussi, bien que l'intentionnalité des traces numériques existe par degré, son inscription est toujours déjà précédée par un cadre d'enregistrement lui-même régi par un modèle d'appréhension des traces sur lequel l'utilisateur n'a pas ou peu prise. Si j'annonce sur Twitter que je suis en train de rédiger ma thèse, mon intention communicationnelle est subsumée par le cadre dans lequel je m'exprime, et dans lequel mon annonce est un tweet, c'est-à-dire un contenu d'un certain format technique et conceptuel qui pourra être appréhendé et manipulé par Twitter ou des tierces parties qui l'interpréteront et en feront un indice d'autre chose. En ce sens, mon tweet est pour moi une trace de mon annonce (avec laquelle elle entretient une relation organique), et pour ces tierces parties une trace d'ordre supérieur, à savoir une trace de mon intention de produire une trace de mon annonce.

Par ailleurs, la question de l'intentionnalité traduit une façon de problématiser la trace numérique qui est centrée sur des agents humains. Lorsque le fonctionnement d'un programme produit des logs comme nous l'évoquions plus haut, il est difficile de parler du degré d'intentionnalité ou de la construction de soi du programme. Les systèmes d'exploitation et les programmes accomplissent une multitude de tâches planifiées ou prévues qui ne sont pas le résultat direct de l'intervention d'un utilisateur ; ils sont en soi des agents non-humains qui peuvent produire des événements traçables. D'un point de vue philosophique, cela ne veut pas dire que les programmes ont des intentions, et on dira plutôt qu'ils sont le résultat de l'intentionnalité du développeur informatique qui les a ainsi conçus, mais cette question de l'intentionnalité des programmes est relativement indifférente pour notre problématique : il suffit de dire que toutes les traces numériques ne sont pas le résultat de

l'action d'un utilisateur. Or si la littérature privilégie les traces produites par des utilisateurs humains, en vertu des considérations anthropologiques qu'elles permettent de thématiser, rien dans la nature de la trace numérique n'impose cette limitation : qu'il s'agisse de sa matérialité technique ou de son interprétabilité, son origine humaine n'apparaît pas comme une distinction pertinente ou nécessaire, ni d'ailleurs discernable nettement. Affirmer d'ailleurs qu'une trace est d'origine humaine ou non humaine est en soi une interprétation. Les systèmes techniques présentent une complexité qui n'est pas un accident mais bien une nécessité de leur fonctionnement, et qui est telle qu'il n'est pas toujours possible de discerner la part de l'un et de l'autre. D'un certain point de vue, toutes les traces numériques sont d'origine humaine dans la mesure où elles proviennent de dispositifs techniques construits par l'homme. D'un autre point de vue, l'utilisation de ces dispositifs ne relève pas pour les humains d'une intentionnalité pure, comme un acte de volonté qui ne rencontrerait ni obstacle ni contrainte, mais d'un geste techniquement instrumenté, et conditionné de ce fait par cette instrumentation.

Par exemple, Twitter est connu pour la limitation de 140 caractères qu'il imposait (jusqu'en novembre 2017) aux messages postés ; ce n'est néanmoins que l'une des innombrables contraintes qui me sont opposées quand je veux écrire un tweet. Le simple fait d'accéder à Twitter me caractérise culturellement, socialement et géographiquement : j'ai de bonnes chances d'être plutôt jeune, plutôt un homme, plutôt urbain, vivant plutôt dans un pays développé, et d'avoir une certaine connaissance et une certaine maîtrise des usages numériques ; j'ai tout intérêt à savoir ce qu'est un *hashtag* ou une *mention*. Je dois disposer d'une machine connectée à Internet, dont je dois utiliser le clavier ; si je veux utiliser des caractères exotiques, n'apparaissant pas sur les touches de mon clavier, je dois savoir comment les produire, ou où les trouver. Twitter peut m'empêcher d'utiliser certains de ces caractères, et peut *a priori*, me bloquer l'accès à ses services, ou *a posteriori*, supprimer un de mes messages. J'utilise peut-être pour rédiger mon message sur mon téléphone portable, un système d'aide à la rédaction par autocomplétion, qui va « corriger » automatiquement mon message, me proposer des termes ; je peux ne faire comme choix d'écriture que d'accepter systématiquement le premier mot proposer par l'outil. Je peux jouer et m'affranchir dans une certaine mesure, de cette limitation de 140 ou 280 caractères, en publiant une image d'un texte, ou un lien vers un texte plus long.

Toujours est-il que je dois m'accommoder de ces contraintes imposées directement par l'outil (ou plus précisément, par les couches de matériel et de logiciel que j'utilise), ou indirectement par les normes et habitudes constituées avec l'utilisation de l'outil. Notre ambition ici n'est pas de faire une monographie complète des prédéterminations de l'utilisation de Twitter, mais de montrer qu'en tant qu'utilisateur d'un système produisant des traces, je suis socio-culturellement et techniquement conditionné dans le message que j'écris, et qui résulte d'un jeu entre mon intention originelle de communication et la forme matérielle qu'elle prendra (ce qui est vrai, d'ailleurs, de tout support d'expression y compris l'oralité).

Par ailleurs, il peut être difficile de retrouver la source de l'intention même à l'origine d'une trace, et de continuer à soutenir son irréductibilité dans le système informatique. Tout ordinateur produit au moment de démarrer des logs qui inscrivent les différents processus exécutés pour son démarrage : où y a-t-il de l'intention dans ce cas ? Peut-on vraiment réduire cet enregistrement à l'action de l'utilisateur d'allumer son ordinateur, ou au travail du développeur matérialisé dans le programme qui produit les logs ? Pour reprendre l'exemple de Twitter, si je configure un robot logiciel qui publie des tweets à ma place, puis que j'oublie son existence et que je le redécouvre quelques temps plus tard, peut-on encore parler d'intentionnalité ? Ou si pour configurer ce robot, j'ai utilisé un outil de création qui me demande seulement de paramétrer un certain nombre d'éléments, mon niveau d'intentionnalité dépend-il du nombre d'éléments configurés ? Si enfin, je n'ai rien à faire, si ce n'est à accepter de fournir mes propres tweets comme référence, à partir de quoi de nouveaux tweets seront générés automatiquement par recombinaison, peut-on encore dire qu'il y a un humain derrière ce compte ? Sachant qu'on estime qu'entre 5% des comptes, ou 24% des messages postés sur Twitter proviennent de *social bots*[16] et que certains sont parfois indiscernables d'utilisateurs humains, est-ce bien légitime de considérer une intentionnalité « directe » *via* une notion d'auteur, comme une caractéristique inhérente à ce réseau social ?

Comme on le voit à travers ces quelques exemples, il n'est pas certain que les notions, culturellement et théoriquement chargées, d'auteur ou d'intentionnalité, soient pertinentes pour caractériser la production de traces numériques ; tout au mieux peut-on, pour sauver cette notion, considérer que l'intentionnalité est distribuée entre les différents agents humains qui ont contribué d'une manière ou d'une autre à l'outil utilisé, et qu'un programme qui fonctionne automatiquement actualise *a posteriori* l'intention de ses concepteurs. En fin de compte on s'aperçoit que la non-intentionnalité est une modalité probablement plus générale, moins susceptible de se heurter à des exceptions, qui s'applique aux agents humains comme aux agents non-humains. Par ailleurs, plutôt qu'une séparation binaire entre ces deux types d'agents, il nous semble qu'en réalité il vaut mieux parler d'agents tous hybridés par de l'humain et du non-humain, suivant des degrés et des modalités qui leur sont propres.

6. Les données comme traces théoriquement chargées

Cette hybridation de l'humain et du non-humain, qui n'a rien d'anecdotique, prive *a priori* les traces numériques de leur statut de représentation de phénomènes humains. Il y a d'une part, des faits humains qui ne produisent pas de trace, et d'autre part, des traces qui ne désignent pas des faits humains. Il faudra, pour regagner ce statut, construire des *données* du fait humain par sélection et construction théorique des traces (par exemple, extraire des tweets dont on serait raisonnablement

[16] https://www.ipsos-mori.com/researchpublications/publications/1760/The-road-to-representivity.aspx (consulté le 16 novembre 2017)

sûrs que les auteurs soient humains), ou à l'inverse, prendre précisément pour objet d'enquête les modalités de cette hybridation.

La trace est pour ainsi dire la donnée à l'état de nature, avant qu'elle n'ait été saisie dans un geste épistémologiquement constitutif et nécessaire aux opérations futures. Ce geste est un choix, une sélection dans « le donné » des traces, orienté par une conceptualisation ou une visée théorique ; ainsi, comme chez Duhem ou Hanson notamment (Heidelberger 2003), les faits sont théoriquement chargés *(theory-laden)*. L'observation des faits n'est pas le fait lui-même, elle est déjà une interprétation motivée par un contexte psychologique, épistémique, mais aussi technique. On peut ainsi comparer les traces numériques aux phénomènes physiques chez Duhem :

> Une expérience de Physique est l'observation précise d'un groupe de phénomènes, accompagnée de l'interprétation de ces phénomènes ; cette interprétation substitue aux données concrètes réellement recueillies par observation des représentations abstraites et symboliques qui leur correspondent en vertu des théories que l'observateur admet. (Duhem 1906)

Suivant notre comparaison, la constitution de données à partir de traces numériques est une expérience (de nature observationnelle plutôt qu'interventionnelle) qui fait de ces données des représentations symboliques. La transaction technique (la collecte et le stockage des données, qui implique une certaine modélisation informatique de celles-ci) s'accompagne d'une transaction épistémique qui leur confère une certaine intelligibilité avant même d'avoir été manipulées, et sont la condition de possibilité de leur intelligibilité ultérieure. L'intelligibilité initiale procède d'une confiance dans la validité des données et d'une reconstitution *a posteriori* de son origine et de ses conditions d'émergence. De ce fait, cette comparaison confirme l'assertion bien connue selon laquelle il n'existe jamais de données brutes, et inscrit toute manipulation de données issues de traces numériques dans un constructivisme qui invalide toute démarche radicalement empiriste. Ce constructivisme présente, on l'a vu, deux aspects principaux :

- un aspect épistémique, comme on l'a vu chez Duhem, par lequel les données deviennent des représentations symboliques ;
- un aspect technique, sur lequel on reviendra au chapitre V, par lequel elles suivent une représentation formelle établie par anticipation, en vue des manipulations que l'on prévoit d'opérer sur elles.

Il intègre aussi un aspect psychologique, qui consiste traditionnellement à penser que nos perceptions des phénomènes sont elles aussi chargées de théorie, voire de préconceptions ou de croyances (Heidelberger 2003). Ici, il nous permet de maintenir une position constructiviste selon laquelle la posture radicalement empirique évoquée ci-dessus n'est pas tant une posture épistémologique que psychologique, ou *a minima* un ethos suggérant des comportements et des croyances. En ces termes, les données sont une préconstruction qu'il convient de déconstruire avant

de les analyser. Autrement dit, cette posture est elle-même une croyance qui conditionne la perception des traces numériques ; bien qu'elle soit un problème pour la construction épistémique de la donnée, elle n'en demeure pas moins une modalité de la collecte de données (par les informaticiens notamment) qui ne peut qu'avoir des conséquences sur les choix théoriques et techniques qui préluderont à cette collecte.

Par ailleurs, dans le geste épistémologique constitutif de la sélection des données, la trace devient également plurielle et perd momentanément son indiciarité. Elle n'est plus l'indice d'un fait singularisé dans un environnement mais un élément parmi d'autres comparables et commensurables dans un ensemble normé : un corpus, un jeu de données. De ce fait, plusieurs types d'analyse, relevant de normes distinctes, sont possibles. D'une part, on peut analyser les traces devenues données de manière quantitative, en recherchant des lois ou quasi-lois mesurables, des moyennes, des constantes statistiques. Les écarts par rapport à ces lois sont des aberrations, du bruit statistique, de même genre que l'erreur de mesure dans les sciences de la nature. Avec ce type d'analyse, les traces numériques rendent possible une appréhension nomologique des phénomènes humains, envisagés selon leurs régularités statistiques, non pas nécessairement mathématisés comme dans le paradigme galiléen, mais quantifiables, manipulables et calculables. On peut par ailleurs reconstituer une indiciarité de la trace, non plus au niveau individuel de l'événement singulier, mais précisément au niveau statistique, en interprétant les écarts aux lois comme des suppléments de sens plutôt que comme des erreurs de mesure. Avec ce type d'analyse, l'approche nomologique est également nécessaire mais peut n'être que transitoire. C'est une étape intermédiaire de la constitution du sens, qui ne détermine pas le type de savoir produit. Ce qui le détermine n'est pas le type de traitement pratiqué, mais ce qui en est fait, la façon dont il est utilisé et interprété.

Par cette pluralité d'approche, les traces numériques placent les sciences computationnelles de la culture dans une troisième voie épistémologique qui n'est pas celle des sciences idiographiques, ni celle des sciences de la nature, et repose sur de nouveaux modes d'accès, en particulier *via* le web, à des phénomènes humains empiriques d'un certain genre. L'épistémologie des traces numériques n'est ni un naturalisme à la Durkheim, qui considère les « faits sociaux comme des choses », bien qu'elle présente des similarités avec les sciences sociales quantitatives durkheimiennes, ni un « culturalisme » fondé sur la compréhension de phénomènes singuliers à partir de l'identification de suppléments de sens dans une norme. C'est une épistémologie qui permet en réalité les deux voies.

Le naturalisme, suivant la définition proposée par Daniel Andler, est « la thèse philosophique selon laquelle toutes les sciences doivent et peuvent viser à traiter de leurs objets respectifs à la manière des sciences de la nature » (Andler 2011). Suivant cette définition, les nouveaux modes d'accès aux phénomènes par les traces numériques ne sont pas ceux des sciences de la nature, et ne conduisent donc pas à une naturalisation des sciences de la culture. Ils maintiennent donc un bifurcationnisme, c'est-à-dire, toujours selon Andler, l'affirmation d'une distinction tranchée d'avec les sciences de la

nature. Néanmoins, la bifurcation imprimée est double : à la fois avec les sciences de la nature, car l'épistémologie des traces numériques n'induit pas un traitement des faits sociaux purement à la manière des sciences de la nature, mais aussi avec les sciences de la culture « traditionnelles », dans la mesure justement où ils se substituent (ou s'ajoutent) aux modes classiques d'accès aux phénomènes humains, et rendent possible à la fois une voie naturalisante et une voie culturalisante.

L'épistémologie de la trace numérique, et de la donnée qui en provient, est une épistémologie des sciences de la culture, dont l'empirie visée est le fait humain, et dont le paradigme indiciaire revisité (car dépossédé de sa dimension idiographique) relève traditionnellement de cette région des sciences. L'exploitation de ces traces numériques repose donc sur une conceptualisation de ces traces qui doit être spécifiquement construite pour acquérir une légitimité scientifique *(a priori*, celle des sciences de la culture), et dépend à la fois de la nature matérielle et sémiotique de ces traces, du projet scientifique poursuivi, et de l'approche adoptée pour les faire se rencontrer.

La conceptualisation des traces n'est pas en effet une opération qui peut se faire dans l'absolu, pour toutes les formes de traces, ou même pour toutes les formes de traces numériques. Nous pouvons même aller plus loin et affirmer que cette conceptualisation est systématiquement faite *ad hoc*, au moment de la constitution des traces en données, même si une première conceptualisation plus générale peut préexister à cette constitution. À ce niveau plus général, nous avons vu qu'il peut y avoir autant de manières d'aborder les traces numériques que d'auteurs les mobilisant, et d'autre part que chaque source (comme Twitter par exemple) présente des spécificités intrinsèques. Cette conceptualisation est à réintégrer dans une continuité méthodologique qui fait également intervenir un projet épistémique, un cadre conceptuel, des outils et méthodes d'analyse et d'interprétation. Nous allons donc interroger l'existence d'une telle continuité, requise pour l'établissement des sciences computationnelles de la culture, et les caractéristiques des composantes qu'elle mobilise.

Chapitre IV.
Vers une continuité épistémique ancrée dans les sciences de la culture

Jusqu'à présent nous avons établi que la donnée numérique possède une valeur épistémique potentielle qui doit pouvoir s'actualiser, et que par son format technique et son caractère quantitatif, elle est également computationnellement manipulable et calculable. Du fait de l'origine et du contenu sémantique des données, issues des traces d'activité humaine (ou présumée humaine), le cadre théorique et méthodologique dans lequel ces traitements calculatoires s'effectuent est de droit celui du paradigme indiciaire, historiquement privilégié par les sciences de la culture, mais applicable à toute donnée conçue comme trace langagière. La donnée joue pour ces sciences le rôle d'un observable qui, afin de générer de nouvelles connaissances, doit être soumis à une analyse qui prend la forme de traitements computationnels. Dans cette configuration *de jure*, la théorie de la donnée et de son traitement est celle de la discipline qui régit l'expérience dans son ensemble : les observables se comprennent au regard de normes disciplinaires qui permettent de les inscrire dans un cadre théorique. Ce cadre fournit ou s'accorde également avec une théorie de l'instrument de sorte que ce sont les mêmes normes qui régissent l'observation, l'analyse et la production de résultats scientifiques. Notre paradigme épistémologique devra lui aussi assurer cette continuité épistémique, en respectant deux contraintes supplémentaires : se fonder sur l'exploitation de traces numériques (par opposition à d'autres observables) et les considérer comme des représentations du fait humain.

Différents projets de recherche plus ou moins féconds dans les sciences de la culture historiquement constituées ont tenté, explicitement ou non, de construire une telle épistémologie de la trace numérique et d'en faire une donnée théoriquement chargée, mais sans réaliser la configuration hypothétique des sciences computationnelles de la culture. Comme nous allons le voir en effet, aucun ne présente toutes les composantes requises ; néanmoins, ils révèlent, dans leurs apports comme dans leurs lacunes, différentes stratégies méthodologiques qui seront exploitées dans notre proposition de paradigme. Concrètement, plusieurs critères sont à examiner pour confirmer l'hypothèse d'une continuité épistémique entre l'observable, les outils, la méthode et la théorie générale :

- l'inscription des observables dans un cadre théorique auquel ils sont reliés à l'aide du paradigme indiciaire tel que nous l'avons conceptualisé pour les sciences de la culture ;
- l'inscription de la méthode d'analyse de ces observables dans ce même paradigme ;
- une compatibilité entre les outils et les méthodes, provenant généralement d'une genèse conjointe des stratégies pour aborder le réel et des outils pour appliquer cette stratégie.

Il faudra également nous assurer que les observables mobilisés sont bien des traces numériques et qu'ils ont un statut de représentation indiciaire du fait humain.

Afin d'identifier des stratégies méthodologiques pertinentes, nous allons examiner ces différents projets de recherche. Nous verrons ainsi les difficultés théoriques et pratiques des sciences de la culture historiquement constituées à s'approprier des objets numériques. Ces difficultés sont notamment examinées au travers de l'exemple des *digital humanities*, dont l'ambition disciplinaire extensive nous permettra de tracer une ligne de partage cruciale au sein des sciences de la culture : les disciplines dont le mode d'accès au réel est le terrain, et celles dont c'est le corpus.

Dans les **disciplines à terrain** d'une part, l'avènement du numérique a permis une transposition des méthodes existantes qui restent fondées sur le paradigme de l'échantillon représentatif. Nous examinons ainsi en détails cette transposition, qui se fait à travers une reconstruction de la continuité épistémologique entre les observables et la théorie, et qui pourra être réexploitée pour notre paradigme méthodologique. Dans les **disciplines à corpus** d'autre part, le numérique induit une révolution technique du texte qui doit être traitée de manière conceptuelle et pratique, par l'acquisition d'un savoir-faire ou la construction d'une interdisciplinarité. Bien qu'elle ne mobilise pas des traces numériques à proprement parler, la statistique textuelle appliquée aux sciences sociales remplit ainsi les critères d'une reconceptualisation du texte par le numérique et de sa mise au service d'une compréhension de faits culturels (et non simplement textuels). Elle manifeste une continuité entre le texte numérique, son analyse et son interprétation qui correspond à ce que nous cherchons à construire à partir des traces numériques.

À ce stade, l'étape manquante est la conceptualisation non pas de données numérisées ou importées, mais de traces nativement numériques, abordées ici à travers l'exemple du web. On verra ainsi qu'il existe plusieurs projets de conceptualisation du web comme trace ou source de traces, qui se sont construites en le considérant d'abord comme un objet en soi, puis comme un intermédiaire indiciaire. De ce fait, aucun projet de recherche au sein des sciences de la culture ne propose de continuité épistémique complète entre les traces numériques, les outils et méthodes et le cadre conceptuel, mais chaque élément analysé en fournit une partie. Et les assemblant, notre paradigme méthodologique peut bénéficier d'une viabilité épistémologique qu'il faudra compléter avec une faisabilité technique, abordée dans le chapitre suivant. Dans un premier temps, nous examinons comment les sciences de la culture mobilisent le numérique d'une part, et ses traces ou données d'autre part.

1. Le numérique dans les sciences de la culture

Nul ne pourrait reprocher aux acteurs classiques des sciences de la culture de ne pas avoir constaté, problématisé et interrogé l'avènement du numérique en général, et du « déluge de données » en particulier, pour leur discipline : il y aurait à opposer à ce reproche l'ensemble des travaux théoriques et empiriques depuis 20 ans sur le numérique et son rôle dans la culture, la société, les médias, la politique, etc.

Cependant, il y a une différence conséquente entre une réflexion intellectuelle *sur* le numérique, et la mobilisation *du* numérique pour la production de connaissances. Cette deuxième option reste un fait rare dans les sciences de la culture en environnement universitaire. Les sciences computationnelles de la culture que nous avons conceptualisées, et qui visent à produire des connaissances à partir du numérique, sont un exemple d'instrumentation du numérique par les sciences de la culture au même titre que l'écriture, les mathématiques, les appareils de mesure physique, instrumentent les sciences de la nature. Néanmoins, les travaux entrepris par les chercheurs en sciences de la culture n'ont dans l'ensemble pas construit une réelle épistémologie de la trace numérique, qui les charge d'une signification ancrée dans un projet épistémique.

Les raisons de ce non-résultat résident vraisemblablement dans le fait qu'une telle épistémologie doit tout d'abord pouvoir se donner un objet, et qu'il doit lui être empiriquement accessible, au risque de l'envisager de manière purement putative et de le manquer : sans confrontation au réel, la réflexion sur un objet fortement médiatisé et politisé court le risque de s'encombrer d'une charge affective qui substitue à l'objet ce que l'on croit ou craint qu'il ne soit. Les épistémologies des différentes disciplines scientifiques historiquement constituées se sont rarement construites *a priori*, comme peut le faire une métaphysique des sciences, mais *a posteriori*, en reflétant les données, les outils et les pratiques d'une époque. Partant que le discours rationnel sur l'exploitation des *big data* peut, lui aussi, difficilement s'établir sans se fonder sur des pratiques effectives de recherche *avec* les *big data*, l'existence et l'accessibilité de ces expérimentations constituent la condition matérielle de possibilité du succès d'un tel discours. Face à la rareté de ces pratiques, la réflexion peine à se les donner comme objet.

Cette rareté des pratiques effectives de ce que nous avons problématisé comme les sciences computationnelles de la culture peut à son tour s'expliquer, en plus des difficultés épistémologiques qu'elles posent, par des limitations techniques, économiques, juridiques et éthiques. D'un point de vue technique, elles requièrent souvent un niveau important de compétences pour fabriquer les données de la recherche d'une part, et les manipuler techniquement d'autre part. Dans le cas des *big data*, la maîtrise de ces enjeux techniques constitue un champ de recherche à part entière, difficilement maîtrisable par les chercheurs en sciences de la culture : comme on l'a vu dans l'analyse de leur mythologie, les *big data* ne consistent, pour beaucoup, que dans la conception et l'utilisation

de ces techniques computationnelles de traitement de données. Cette technicité pose des enjeux d'interdisciplinarité spécifiques que nous abordons par la suite.

Par ailleurs, la configuration d'acteurs que nous avons déjà évoquée plusieurs fois pose des freins économiques et juridiques à de telles pratiques. Les données et les outils sont bien souvent la propriété d'entreprises qui en monnaient l'accès ou ne le proposent tout simplement pas. Dans ces configurations, l'analyse de données massives est tout simplement réservée aux entreprises ; quand bien même les chercheurs auraient le désir, et l'appareil théorique et pratique, pour traiter de telles données, ils ne pourraient techniquement pas y accéder – ou difficilement. La constitution de données et d'outils requiert elle aussi des compétences et des moyens matériels dont ne disposent pas forcément, voire jamais, les chercheurs en sciences de la culture. Enfin, certaines données accessibles techniquement sont soumises à des problématiques juridiques ou déontologiques telles que la protection de la vie privée, le traitement de données personnelles. Là où les sciences de la nature s'efforcent depuis quelques années de lever les freins à la recherche posés par la propriété des données, notamment par l'intermédiaire d'initiatives d'ouverture et de partage de données, les sciences de la culture ont en commun avec le monde médical le caractère sensible de la donnée elle-même, indépendamment de la propriété de celui qui les a établies ou collectées. L'exploitation de données personnelles à des fins de recherche met alors en tension deux exigences juridiques et éthiques : celle du respect de la vie privée (et de la propriété intellectuelle) d'une part, et celle de fournir à la recherche publique les moyens nécessaires à la poursuite de ses travaux. Le projet de loi « pour une République numérique » va dans le sens de la seconde exigence en autorisant la reproduction de données numériques par les chercheurs dans le contexte de « l'exploration de textes et de données pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale »[17]. Les enjeux éthiques relatifs aux données personnelles, à la question du consentement, et aux usages potentiellement problématiques des données (et des traitements) demeurent néanmoins d'actualité malgré des assouplissements juridiques en faveur des chercheurs.

Indépendamment de ces limitations techniques, économiques, juridiques et éthiques, bon nombre de travaux de recherche en sciences de la culture relatifs à l'analyse de données numériques sont davantage de nature réflexive et critique que de nature expérimentale. Il existe un foisonnement de champs de recherche sur le numérique développés dans ce contexte, tels que les *digital humanities*, les *digital studies*, les *cultural studies*, les *software studies*, les *critical data studies*, et bien d'autres encore, qui ont en commun de se construire davantage comme une réflexion prospective et critique sur le numérique et la donnée que comme un travail effectif avec ces éléments. Les *digital humanities* constituent à ce titre l'exemple emblématique de ces projets chargés en imaginaire, qui produisent principalement des discours sur les techniques sans usages techniques, occasionnant un décalage

[17] Projet de loi « pour une République numérique », amendement 180, http://www.assemblee-nationale.fr/14/amendements/3399/AN/180.asp (consulté le 16 novembre 2017)

entre leurs ambitions et les moyens qu'elles se donnent. Elles sont pour ainsi dire le pendant ancré dans les sciences de la culture de la mythologie plus générale des *big data*. Nous allons en examiner les ambitions, la structuration et le rapport à la technique.

2. Terrain et corpus : deux modes d'accès au réel

Les *digital humanities* sont habituellement décrites comme un projet relatif à l'articulation entre les *humanités* et le *numérique*, tous deux entendus dans un sens très large. En français comme en anglais, la notion de *digital humanities* a davantage une fonction programmatique qu'une fonction descriptive. Les documents emblématiques du sujet sont des *manifestes*, ce qui suggère qu'il s'agit plus de l'élaboration d'un programme ou d'un ensemble d'ambitions pour les chercheurs en « humanités » que d'une notion analytique avec laquelle on analyserait rétrospectivement un phénomène dans une approche sociologique ou épistémologique. Ledit programme des humanités numériques porte à la fois sur la compréhension des enjeux du numérique, la mobilisation des technologies numériques pour la recherche, mais aussi des enjeux plus larges d'organisation, de collaboration et d'interdisciplinarité entre chercheurs et que les dispositifs de *digital humanities* devraient rendre possible. Ce programme est aujourd'hui largement remis en cause sous tous ces aspects, et son caractère spéculatif bien connu et critiqué. Les travaux sur les humanités numériques sont de plus en plus des critiques de celles-ci, qui en dénoncent notamment les motivations politiques et idéologiques. Pierre Mounier, qui est par ailleurs un acteur reconnu du domaine, dénonce ainsi le caractère « d'utopie politique » (Mounier 2015) d'un tel projet, ainsi que le décalage entre les ambitions de cette utopie et les moyens et forces en présence. Eric Guichard (2013) décrie plus largement la rhétorique d'une époque « saturée de discours révolutionnaires » largement idéologiques ; on pourrait également pointer les disproportions entre les ambitions des humanités numériques (Berry 2012; Burdick et al. 2012) et les transformations concrètes des disciplines concernées. Les rédacteurs du manifeste anglophone conviennent eux-mêmes du fait qu'il ne faut pas s'appesantir sur ce mot d'ordre dont l'emploi correspond à une tactique d'appropriation de la problématique :

> We wave the banner of « Digital Humanities » for tactical reasons (think of it as « strategic essentialism »), not out of a conviction that the phrase adequately describes the tectonic shifts embraced in this document. But an emerging transdisciplinary domain without a name runs the risk of finding itself defined less by advocates than by critics and opponents, much as cubism became the label associated with the pictorial experiments of Picasso, Braque, and Gris. (*Digital Humanities Manifesto 2.0*)

Cette tactique, comme d'ailleurs la rhétorique des *big data*, sert notamment des enjeux de notoriété et de financement de la recherche ; il joue également un rôle de reconnaissance et de ralliement entre chercheurs venant de champs différents et intéressés par une problématique commune. L'imprécision et la désorganisation apparente du projet n'est pas considérée comme une faiblesse conceptuelle mais comme une « force opérationnelle » (Mounier 2015) qui permet l'échange et l'interdisciplinarité.

Néanmoins, il est difficile de définir quelle est cette problématique commune, tant les *digital humanities* peuvent concerner à la fois la réflexion sur le numérique, l'utilisation d'outils numériques, le rôle du numérique pour les « humanités », et de manière générale toute articulation entre les termes d'humanités et de numérique. En particulier, les humanités numériques françaises sont marquées par une forte hétérogénéité en matière de champ disciplinaire et donc de culture épistémique, car elles intègrent, d'après le manifeste francophone des *digital humanities* « l'ensemble des Sciences humaines et sociales, des Arts et des Lettres »[18]. Cela signifie qu'elles considèrent que l'on peut faire entrer dans une même problématique :

- d'un côté les *humanities* du monde anglo-saxon, c'est-à-dire des disciplines qui accèdent au réel par le livre et l'archive, ou dont l'objet de recherche même est précisément l'un de ces supports (comme par exemple les *media studies*) et qui s'intéressent donc tout particulièrement aux mutations de ces objets avec le numérique ;
- et de l'autre côté les sciences humaines et surtout sociales, dont le mode d'accès au réel, qualitatif ou quantitatif, est principalement l'observation de l'événement, et non sa trace.

Concrètement, elles mobilisent un ensemble de disciplines historiquement constituées que l'on peut répartir selon leur mode d'accès au réel, avec d'un côté les disciplines centrées sur l'étude de textes et d'archives, et de l'autre celles qui s'appuient sur une recherche de terrain (**Tableau 4**). Ces familles de disciplines se distinguent par le niveau de médiation et donc de formalisation de leur rapport au réel : dans les **disciplines à corpus**, le réel est toujours déjà médiatisé par un support qui apporte avec lui sa propre grammatisation, sa forme et son format, et constitue donc toujours déjà une transformation de l'objet étudié. Ces disciplines sont fortement ancrées dans le paradigme indiciaire. On n'étudie jamais l'événement historique, mais un exemplaire matériel d'un texte qui est lui-même une trace de l'événement. Ce format prescrit des usages et manipulations possibles au sein d'une culture lectrice et interprétative. Dans les **disciplines à terrain** au contraire, le réel doit être saisi, capturé, et traduit sur un support constitué *ad hoc* pour l'analyse et l'interprétation. Pour l'ethnographie, l'observation du phénomène doit être formulée, inscrite et grammatisée par le chercheur pour en constituer le sens, soit par une prise de note pendant l'observation, soit par la conduite de l'entretien qui amène le témoin à une formulation verbale, linguistique, de son expérience du phénomène (qui peut d'ailleurs être précisément le phénomène étudié lui-même). Dans les méthodes quantitatives, peut-être moins linguistiques et plus formalisées, le chercheur doit ici encore produire ses observables, régis par des normes statistiques. S'il existe ainsi plusieurs techniques normées d'observation, d'entretien et de sondage, elles ne permettent pas de mécaniser la traduction des observations, qui relève toujours d'un savoir-faire propre à chaque chercheur. Ainsi, dans les

[18] http://tcp.hypotheses.org/318 (consulté le 20 novembre 2016)

disciplines à terrain, le chercheur doit toujours *construire* les données là, alors qu'elles doivent être *trouvées* et délimitées dans les disciplines à corpus.

Corpus	**Terrain**
Histoire et archéologie	Sociologie
Arts	Économie
Langue et littérature	Géographie
Droit	Démographie
Sciences de l'information et de la communication	Psychologie
	Anthropologie

***Tableau 4**. Quelques exemples de disciplines classées d'après leur mode d'accès au réel.*

La ligne de partage que nous traçons ici, et les réalités visées de part et d'autre, sont plus floues et complexes que ce que nous avons esquissé, et présentent bien sûr de multiples exceptions. Au sein des sciences humaines et sociales par exemple, le champ des sciences cognitives recouvre à la fois des pratiques entièrement naturalisées, imprégnées de l'impérialisme des sciences de la nature, telles que la psychologie expérimentale, et des champs de réflexion intellectuelle comme la psychanalyse, qui s'apparentent bien plus, d'un point de vue épistémologique, aux travaux critiques des humanités qu'à la recherche d'objectivité des sciences sociales. Cette distinction permet néanmoins de mettre au jour une hétérogénéité au sein des sciences de la culture qui renvoie à la fois à des modes d'accès au réel distincts et à des rapports spécifiques à l'instrumentation des disciplines, et notamment, comme on va le voir, à l'instrumentation du numérique.

Selon ce mode d'accès au réel, l'avènement du numérique se fait en effet différemment, soit qu'il transforme les observables traditionnels des disciplines, soit qu'il leur apporte au contraire de nouveaux observables ou du moins, de nouvelles médiations pour l'observation des phénomènes. De ce point de vue, les humanités numériques francophones passent sous silence la complexité des transformations des sciences de la culture par le numérique, et ne précisent pas quel tout ou partie de ces transformations est visé par elles.

Les racines anglo-saxonnes des *digital humanities* sont moins ambigües : il s'agit bien souvent de projets de numérisation avec des problématiques de philologie numérique. L'exemple fondateur est le *Corpus Thomisticum* de Roberto Busa (numérisation des textes de Thomas d'Aquin) et l'un des plus vivaces est le projet de la TEI *(text encoding initiative)* qui vise à promouvoir un format pour les documents numériques qui intègre les aspects problématiques de l'établissement des textes. Les *digital humanities* anglo-saxonnes relèvent ainsi résolument des disciplines à corpus, et travaillent un rapport au texte d'autant plus crucial qu'il est leur mode d'accès au réel.

Cette question du champ disciplinaire est déterminante, car elle touche à une ligne de partage majeure à l'intérieur des sciences de la culture. En effet, les disciplines à terrain adhèrent assez largement à l'idée d'une certaine naturalisation de leur objet, à partir de laquelle la construction d'une certaine objectivité devient possible. Comme on l'a vu, la sociologie durkheimienne pose les faits sociaux comme des choses et n'est pas antinaturaliste. Sans nier forcément la singularité de leur objet, et

notamment la réflexivité intrinsèque qui la caractérise et que nous avons évoquée au chapitre précédent, les disciplines à terrain considèrent leur pratique comme scientifique (au sens des sciences de la nature), et s'efforcent d'évaluer le travail au prisme d'exigences de rationalité et de critères de validité propres à leur discipline. De là, découle un organon matériel et conceptuel qui permet de se donner des observables, de mesurer des phénomènes, bref de construire un accès au réel cohérent et valide à partir duquel on peut explorer et valider des hypothèses. De ce point de vue, le positionnement des disciplines à terrain relève, comme on l'a vu au chapitre précédent, d'un bifurcationnisme épistémologique qui refuse une réduction des sciences de la culture aux sciences de la nature (et notamment à la physique) mais entend bien néanmoins se proposer comme science combinant une certaine naturalisation de ses objets avec un recul critique et réflexif.

À l'inverse, le positionnement revendiqué par les *humanities* anglo-saxonnes, bon nombre de *critical studies*, et des travaux d'obédience postmoderniste, refusent tout impérialisme scientifique (Dupré 2002) en provenance des sciences de la nature et se présentent comme une investigation intellectuelle autonome vis-à-vis d'elles. Leur ambition n'est nullement de construire une objectivité mathématique ou computationnelle, de dégager des lois ou des quasi-lois, mais de rendre compte de la singularité des objets qu'elles se donnent. De ce fait, elles s'inscrivent souvent dans une épistémologie où le pluralisme des interprétations possibles d'un texte fait partie des donnés de leur discipline.

Disciplines à terrain et disciplines à corpus articulent différents modes d'accès au réel, distinctement transformés par le numérique. Les épistémologies qui en découlent sont également à examiner de manière séparée, bien que des parallèles puissent être tracés. Afin de mettre en évidence leur rapport spécifique au numérique, seulement esquissé jusqu'à maintenant, nous allons maintenant examiner plus en détail le mode d'accès au réel apporté par le numérique tel qu'il est conceptualisé par ces deux familles de disciplines : du côté des disciplines à terrain, nous examinons ce que Richard Rogers (2008) a baptisé « virtual methods », c'est-à-dire la transposition des méthodes traditionnelles d'enquête qualitative et quantitative, et du côté des disciplines à corpus, plusieurs efforts de conceptualisation du web comme corpus ou source de corpus.

3. Les *virtual methods*, une reconstruction de l'accès au terrain

Dans les disciplines à terrain, c'est le web qui est jugé pertinent depuis une vingtaine d'année pour mesurer des phénomènes sociaux et des mouvements d'opinion à travers des enquêtes quantitatives ; il constitue pour elles un mode supplémentaire d'accès au réel, qui s'ajoute, voire se substitue, à des canaux plus traditionnels. Le sondage dit « électronique » est devenu dominant[19] dans les méthodes

[19] Dans les instituts de sondage, les trois quarts des enquêtes quantitatives ont été réalisées par Internet en 2015, d'après une enquête réalisé par le Syntec Etudes, en ligne : http://www.syntec-etudes.com/fichiers/20170320182258_Marche_francais_des_Etudes_2015.pdf (consulté le 30 novembre 2017).
En ligne : http://www.syntec-etudes.com/fichiers/20170320182258_Marche_francais_des_Etudes_2015.pdf (consulté le 20 avril 2017)

quantitatives de recueil de l'opinion pour son caractère presque immédiat, les manipulations qu'il permet et son intérêt économique (il coûte beaucoup moins cher à produire que ses alternatives). Le support numérique est utilisé à deux titres :

- différents canaux accessibles *via* Internet, comme le mail ou les réseaux sociaux, peuvent être utilisés pour recruter des répondants ;
- le web à proprement parler sert de support au questionnaire lui-même.

Le sondage en ligne ne procède pas du développement de nouvelles méthodes ni d'une nouvelle épistémologie : ce sont des *virtual methods* (Rogers 2009) qui transposent des pratiques existantes. Il ne repose pas sur une épistémologie de la trace mais sur une approche nomothétique des phénomènes sociaux, naturalisés par la quantification. Le sondage se fonde sur une épistémologie de la représentativité que l'on retrouve quel que soit le mode de recueil : téléphone, face à face, et maintenant « online ».

Avec l'introduction des sondages en ligne, les disciplines à terrain ont besoin de montrer qu'il y a une continuité avec les modes de recueil, et que la norme de la représentativité est préservée. Les praticiens doivent ainsi pouvoir démontrer la capacité d'un échantillon d'individus à présenter les mêmes caractéristiques que la population à partir de laquelle il est échantillonné. Néanmoins, l'échantillonnage comme critère de validité de la donnée est soumis à un certain nombre de biais liés à son principe même, dépendant ou non du mode d'administration du questionnaire. Ainsi, d'une part la norme des disciplines à terrain n'est pas toujours respectée, et d'autre part, l'introduction du sondage en ligne implique de s'assurer qu'il présente *a minima* les mêmes biais. Ces biais, bien connus des méthodes quantitatives en sciences sociales, réduisent la représentativité des données, et nuisent donc à leur valeur épistémique en amont de l'analyse qui leur est faite.

On peut les regrouper en trois types :

- **Les biais relatifs au mode d'échantillonnage**. En l'absence d'une base de données complète de la population française (dans le cas d'une enquête visant le marché français), il n'est pas possible de sélectionner aléatoirement un échantillon de la population totale. La méthode des quotas, largement utilisée par les instituts de sondages, introduit des biais d'après l'orthodoxie méthodologique. En théorie, seul un échantillonnage aléatoire peut être considéré comme représentatif ; en pratique, les échantillons vraiment aléatoires d'un point de vue statistique sont difficiles, voire impossibles à constituer. De ce fait, la méthode des quotas constitue un pis-aller très largement utilisé et généralement considéré comme suffisant ; elle est de fait jugée plus fiable que les échantillons non représentatifs auxquels les chercheurs en sciences sociales ont recours faute d'un accès à des panels d'individus constitués par quotas comme ceux dont disposent les instituts de sondages.

- **Les biais dus au taux de non-réponse.** Le taux de non-réponse décrit la proportion d'individus qui, soumis au questionnaire, ne l'ont pas rempli. Les raisons de ce choix peuvent être diverses (la personne estime qu'elle n'a rien à dire ou rien à reprocher ; elle présente une incapacité physique ou mentale qui l'empêche d'accéder au questionnaire ou d'y répondre ; elle est opposée au principe du sondage en général ; elle est mécontente de la prestation de l'entreprise et ne souhaite pas lui apporter d'information ; etc.) mais la population des répondants et des non-répondants diffère sous plusieurs aspects. Les motifs de non-réponse peuvent d'ailleurs être précisément le genre d'information que l'entreprise cherche à identifier. Par ailleurs, les questions ouvertes sont parfois facultatives dans les questionnaires, ce qui accroît encore le taux de non-réponses sur ces questions.
- **Les biais dus au mode d'administration du questionnaire**. Les caractéristiques des répondants et la teneur de leurs réponses varient selon le mode d'administration du questionnaire. La littérature universitaire sur ce sujet (un peu vieillissante car souvent antérieure au développement du web, ou contemporaine de la pénétration de l'informatique dans les foyers occidentaux) s'attache plutôt à montrer les différences entre les modes de recueil que la supériorité de l'un sur l'autre. Ainsi les avantages et les inconvénients des différents canaux sont également bien connus :
  - Le **face-à-face** est particulièrement coûteux notamment lorsque l'enquête est réalisée à plusieurs heures et endroits différents, ou qu'elle se fait en salle et non dans la rue (situation inconfortable pour le sondé, surtout pour les thématiques plus délicates). La coprésence physique permet une richesse dans l'échange, pas seulement verbal ; cependant cette richesse se perd dans la transcription de l'oral vers l'écrit, qui transforme le propos des enquêtés. L'enquêteur dispose d'une perception des répondants et de leurs réponses meilleure que celle de l'analyste, du fait des données qui disparaissent dans le mode formel du questionnaire : hésitations, explications, reformulations, communication non verbale, etc. (Lallich-Boidin 2001). Par ailleurs, les enquêtes de terrain montrent un décalage notable entre l'orthodoxie méthodologique attendue de l'administration des questionnaires (aussi bien en face à face que par téléphone) et les pratiques réelles ; la reformulation des questions, la constitution d'une relation personnalisée, sont autant d'éléments théoriquement destructeurs de la scientificité des méthodes quantitatives, mais qui peuvent en réalité la produire en conduisant le sondé à proposer des réponses qui ont du sens. De plus, le face-à-face est connu pour créer de la désirabilité sociale, c'est-à-dire la tendance du sondé à donner des réponses qui donnent une bonne image de lui, ou plus conforme à ce qu'il croit être la norme sociale (Frippiat et Marquis 2010). Enfin, le terrain des sondages réalisés en instituts est la plupart du temps réalisé par des enquêteurs en situation précaire, souvent prestataires de l'institut et non-salariés, et

dont les contraintes de délais font baisser la qualité de l'administration du questionnaire, jusqu'à en inventer les réponses (Marc 2001; Caveng 2012).

- Le **téléphone** permet théoriquement d'utiliser la seule méthode d'échantillonnage considérée comme vraiment valable, à savoir le tirage aléatoire, qui se fait sur les numéros de téléphone. Il a, dans une certaine mesure, les mêmes avantages et inconvénients que le face-à-face du point de vue de l'oralité, mais l'absence de coprésence physique réduit les effets de personnalisation de l'échange et ses conséquences ; le relatif anonymat du téléphone permettrait ainsi de réduire la désirabilité sociale. Néanmoins le taux de non-réponse, en augmentation constante, biaise l'échantillonnage, sur une population de moins en moins représentative notamment du fait du déclin du téléphone fixe, l'utilisation du mobile posant d'autres problèmes (Witte 2009).
- **Internet** est particulièrement ambivalent du point de vue des populations qu'il permet de toucher. Si l'équipement des foyers n'est pas complet puisqu'il était de 65% en 2009 dans l'Europe des 28 (Frippiat et Marquis 2010), le taux continue à progresser (il atteint 85% en 2016 en France[20]) et ce média permet d'atteindre des populations auparavant inaccessibles : des réfractaires au téléphone mais aussi des groupes d'individus unis par une thématique commune ou des caractéristiques partagées. Les usagers d'Internet ne sont pas répartis uniformément d'un point de vue sociologique, mais contrairement à l'intuition, les populations plus âgées, et non les jeunes, peuvent être surreprésentés (Loosveldt et Sonck 2008). Le trait le plus marquant de ce mode de recueil est cependant le biais d'auto-sélection, dû au fait que les questionnaires sont diffusés largement (par mail, sur les réseaux sociaux) à des panels où les sondés choisissent (ou non) de s'inscrire puis de répondre, vraisemblablement en fonction de leur intérêt ou de leur implication éthique ou politique pour la problématique. À l'inverse, Internet permet une meilleure maîtrise du recueil lui-même, avec les outils de mesure et d'analyse des comportements propres au web. Il faut noter enfin que la notion de sondage par Internet fait écran à un réseau complexe de modalités de recueil (**Figure 1**) qui génèrent des niveaux de représentativité variables. Partagé entre ses avantages et ses inconvénients, il n'est finalement rien d'autre que « des potentialités techniques qui augmentent le contrôle du chercheur sur certains aspects et l'amenuisent sur d'autres » (Frippiat et Marquis 2010).

[20] D'après le baromètre du numérique de l'Arcep 2016, http://www.arcep.fr/uploads/tx_gspublication/presentation-barometre-du-numerique-291116.pdf (consulté le 14 avril 2017)

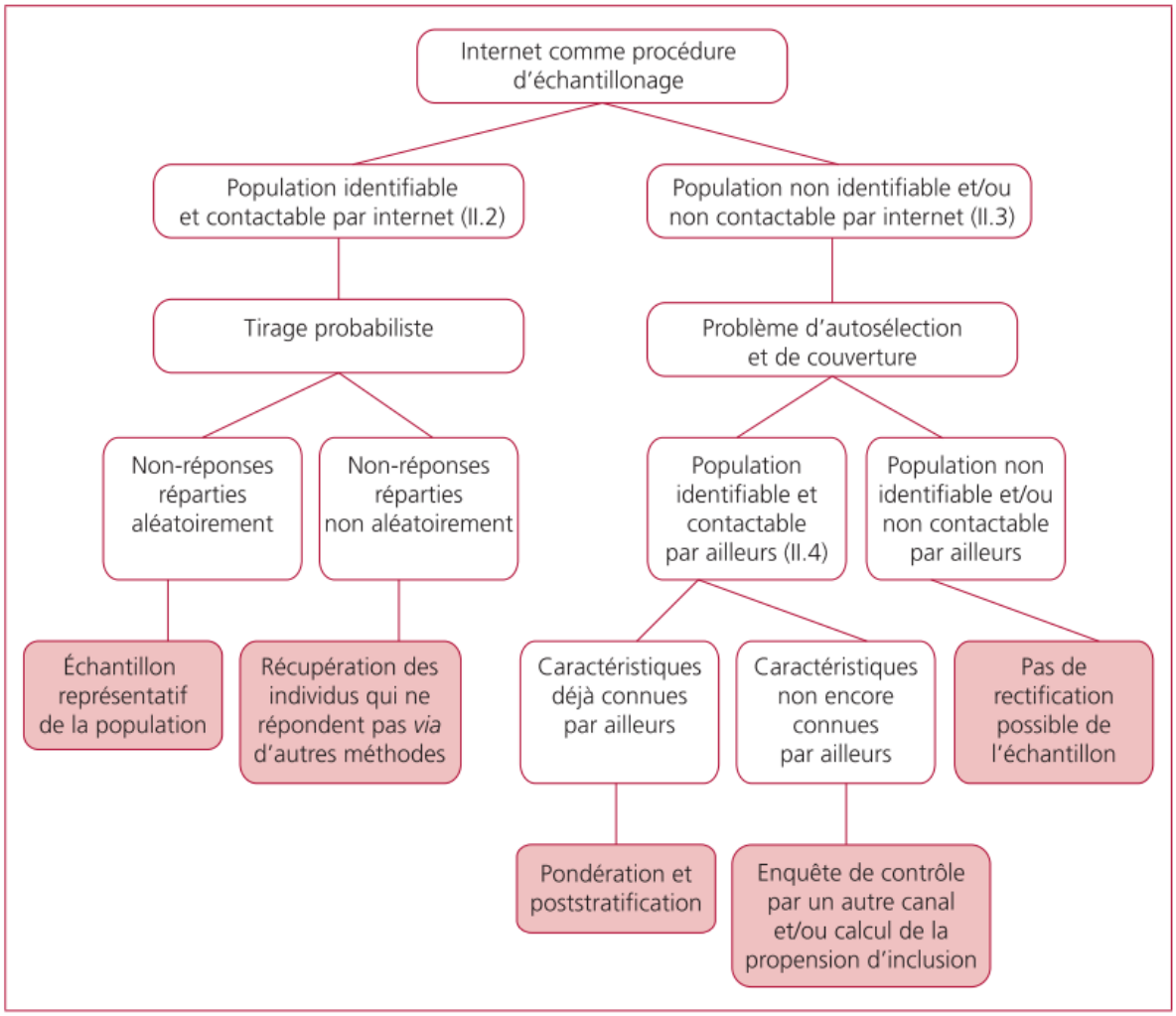

Figure 1. *Différentes modalités d'échantillonnage dans les sondages en ligne.*

Tiré de Frippiat et Marquis 2009

Le recueil par Internet remobilise le support écrit, comme dans les enquêtes postales plus ou moins tombées en désuétude, et constitue ainsi une rupture avec la précédente innovation de l'histoire des sondages, le téléphone : l'histoire des modes de recueil n'est pas linéaire. Que ces méthodes soient commensurables ne veut pas dire cependant qu'elles soient équivalentes. Comme on l'a vu, elles présentent chacune leurs biais spécifiques. Le développement des études d'opinion par Internet a d'ailleurs suscité une littérature assez riche et récente sur la méthodologie des sondages en général. L'apparition d'un nouveau canal d'enquête permet de jeter un regard rétrospectif sur les modes précédents, avec un recul nouveau. Plutôt que d'opposer le nouveau mode aux modes traditionnels, cette mobilisation comparative permet d'intégrer Internet dans les canaux de recueil de l'opinion et d'assurer une continuité méthodologique avec l'existant. La comparaison est possible sur un mode horizontal (dégager les spécificités de chaque mode, plutôt que de comparer à un mode de référence) précisément du fait de l'absence d'un mode de référence à l'aune duquel les autres seraient évalués. S'il y a une orthodoxie méthodologique en théorie, aucun des modes ne l'applique effectivement en pratique. La possibilité de procéder à un échantillonnage effectivement aléatoire *via* le principe du

RDD *(random digit dialing)* a pu donner un certain prestige au sondage par téléphone, utilisé comme référence dans un comparatif avec un sondage en ligne réalisé en 2004 (Schonlau et al. 2004). Dans cette étude, les données collectées en ligne ont été redressées en utilisant les résultats du sondage fait par téléphone, ce qui en fait *de facto* une mesure de référence. On pourrait néanmoins en dire autant du face-à-face qui a lui aussi servi d'élément de comparaison par rapport à Internet (Loosveldt et Sonck 2008).

D'autres comparaisons sont possibles. Leur intérêt est qu'elles permettent aux sondeurs d'expliciter les éléments sur lesquels s'effectue la comparaison, de définir des invariants méthodologiques et des parties modulables. L'un de ces invariants est l'objet lui-même : la mise en comparaison de deux méthodes du point de vue de leur capacité à mesurer un même phénomène est un postulat implicite de l'identité de l'objet mesuré par ces différentes méthodes. D'un point de vue statistique, les échantillons constitués peuvent être comparés de manière chiffrée, soit entre eux, soit par rapport à une population de référence (si celle-ci est accessible). L'enjeu est alors de neutraliser les biais liés au mode de recueil par un calcul qui rend la commensuration effective. Dans la pratique, on peut citer l'exemple du score de propension (Schonlau et al. 2004) qui permet de comparer un premier échantillon à un échantillon de référence ; l'objectif de cette comparaison est de pondérer *a posteriori* le premier échantillon. Dans l'étude en question, l'échantillon de référence a été obtenu par RDD, puis rééchantillonné avec des quotas d'origine ethnique (c'est une étude étatsunienne, permettant légalement d'inclure ce type de variable dans un questionnaire), d'âge, de niveau de pauvreté, etc. Pour effectuer la commensuration, il est nécessaire que les deux sondages aient des questions communes, constitutives d'une même mesure ; en fonction des réponses données à ces questions, chaque individu est doté d'une nouvelle variable, obtenue par calcul, qui permet de redresser l'échantillon premier avec l'échantillon de référence. On est donc là dans un mode de comparaison qui n'utilise qu'indirectement la population de référence et ses caractéristiques connues. La représentativité supposée de l'échantillon de référence, dont on a une connaissance probabiliste, sert d'intermédiaire à la construction de la représentativité statistique de l'échantillon où elle n'est pas connue. En simplifiant un peu le processus, on peut dire qu'il correspond au raisonnement suivant :

- l'échantillon de référence est représentatif de la population générale (postulat de départ) ;
- l'échantillon de référence se caractérise par un ensemble de propriétés chiffrées E ;
- l'échantillon nouveau se caractérise par un ensemble de propriétés chiffrées F ;
- on calcule la différence entre E et F puis on la neutralise en calculant des scores de pondération à appliquer à F pour que les deux échantillons aient les mêmes propriétés ;
- alors l'échantillon nouveau est représentatif de la population.

Ce raisonnement n'est pas une déduction au sens d'inférence du général au particulier, ou dans les termes de la statistique classique, de la population à l'échantillon ; c'est une inférence, à partir d'un

particulier institué comme représentant du général, vers un autre particulier que l'on cherche à investir des mêmes propriétés représentatives.

Pour certains chercheurs, l'économie d'un rapport inférentiel à la population générale n'est pas admissible en théorie ; en pratique, l'accès à cette population peut être difficile ou impossible. Il est rare qu'il existe un registre général, non biaisé, des individus. En France, les chiffres de l'INSEE font souvent référence, mais leur rapport à la population générale est lui aussi une construction, largement analysée en sociologie de la quantification à la suite des travaux d'Alain Desrosières sur le référencement national (Desrosières 2010). La population de référence, qui n'est pas la population générale, est une convention par rapport à laquelle des enquêtes peuvent se situer et se comparer entre elles. L'intérêt de la population de référence est qu'on en connaît plus facilement les caractéristiques à travers des mesures chiffrées qui servent d'étalon pour les échantillons constitués, et à partir desquelles on peut redresser ces échantillons. La nécessité de connaître les caractéristiques de la population de référence n'est pas absolue pour tous les chercheurs : d'un côté, Frippiat et Marquis (2010) affirment que fonctionner sans population de référence « devrait donc signifier que l'on se condamne à abandonner toute velléité de représentativité » tandis que Witte et Howard (1999), cités par Couper (2000), écrivent :

> while the survey did not yield a random sample and the selection probabilities are unknown, "this does not mean that the survey cannot yield representative social science data".

Cette dernière citation met en évidence un point important : si la connaissance de la population de référence est *a priori* souhaitable, elle n'est cependant pas toujours possible.

De plus, cette connaissance n'est pas binaire : on peut disposer de certaines informations relatives à la population générale mais pas de tout. En utilisant ces informations, il est possible de redresser l'échantillon à partir de la comparaison entre la population générale et cette population de référence (et non d'un échantillon de référence comme on l'a vu plus haut). Néanmoins les techniques de redressement, plus ou moins sophistiquées, font également l'objet de critiques, notamment dans le contexte des sondages par Internet : s'il est possible d'ajuster les proportions (sur le principe des quotas) en fonction de réponses ou d'autres variables, cela ne veut pas dire pour autant que l'échantillon sera plus représentatif ni mieux comparable à la population générale (Loosveldt et Sonck 2008). Par ailleurs, les différences d'un mode de recueil à l'autre ne se mesurent pas seulement en termes de représentativité statistique à la population générale ; les enquêteurs s'inquiètent aussi de la qualité des réponses et de la propension des internautes à « bien » répondre au questionnaire, c'est-à-dire sans exagérer ni atténuer leurs réponses (Frippiat et Marquis 2010).

La question de la représentativité reste au cœur des problèmes épistémologiques des sondeurs. Elle traduit une conception des phénomènes sociaux qui présente deux caractéristiques intéressantes :

- D'une part, le sondage est une représentation statistique du réel, fondée sur le nombre et l'approximation. L'individu n'existe qu'en tant qu'il est partie d'un tout où il incarne certaines caractéristiques ; on se situe donc dans la recherche de lois (ou quasi-lois) générales plutôt que dans un effort de caractérisation de ce qui fait la spécificité et l'unicité d'un individu. En revanche, si l'approche est plutôt nomothétique, l'analyse de la variance et d'autres approches statistiques de l'étude des écarts à la norme mettent plutôt en évidence les spécificités d'une sous-population par rapport à la norme, dans une démarche plus idiographique.
- D'autre part, si cette représentation est statistique, elle n'en demeure pas moins conçue comme une correspondance au réel, avec lequel elle partage certains traits mais pas la totalité, avec le problème de déterminer si ces traits sont suffisants pour traiter la question de recherche que l'on se pose. Ainsi le sondeur peut très bien savoir que son échantillon n'est pas homogène à la population générale en termes d'âge ou de genre, mais l'utiliser néanmoins en connaissance de cause, dans la mesure où il étudie un phénomène sur lequel il sait par ailleurs que ces paramètres n'ont pas d'impact.

Les méthodes quantitatives en sociologie héritent au moins en partie leur théorie de la connaissance des techniques mathématiques qu'elles mobilisent : la conception de la vérité comme correspondance statistique n'est pas liée à l'objet ou au champ disciplinaire mais à la façon de l'appréhender techniquement. Elle est à l'œuvre ici de la même façon qu'elle peut l'être dans d'autres domaines d'utilisation plus proches des sciences de la nature, notamment en biostatistique (qui est d'ailleurs le domaine d'invention du score de propension). On aurait ainsi une épistémologie transdisciplinaire organisée autour des concepts d'individu et de population, qui sont bien dans les faits communs à la biologie et à la sociologie quantitative. La statistique ne jouerait pas un rôle purement instrumental, où elle ne serait qu'une technique au service des sciences, mais constituerait bien par elle-même une manière de former des savoirs, tant d'un point de vue opérationnel que théorique.

Toutefois, bien que le web constitue ici un nouveau mode de mesure des phénomènes sociaux, ses spécificités n'y sont mobilisées que de manière très superficielle, et ce dans le but d'assurer une continuité avec des pratiques existantes. À ce titre, il apparaît à la fois comme une opportunité pratique (il coûte moins cher, permet de mieux maîtriser les modalités de collecte) et comme une réponse à l'effritement de la représentativité des sondages, dont le taux de réponse n'a cessé de diminuer depuis Gallup. Dans ce contexte, il n'est cependant qu'un médium, et non une source. Les sondages en ligne ne produisent pas des traces numériques mais des données relatives aux individus par l'intermédiaire du web.

Ces *virtual methods* présentent cependant un certain nombre d'enseignements pour la conceptualisation des traces numériques, notamment en termes de construction d'une continuité épistémologique avec des méthodes antérieures. On voit en effet qu'il y a une certaine abondance de travaux dédiés à l'évaluation de cette continuité entre mode de recueils, et à l'élaboration de critères

de comparaison. Ainsi, le changement de support par exemple (écrit vs oral) n'est pas surmonté par une démonstration de la commensurabilité de l'écrit et de l'oral, mais par une mise en comparaison des résultats produits respectivement par l'un et l'autre. Les *virtual methods* se sont dotées d'outils comme le score de propension pour comparer les représentations du réel que produisent les différents modes de recueil : cette comparaison se fait *a posteriori*, sur des exemples empiriques, plutôt qu'en discutant les principes respectifs des différentes approches. La norme de la représentativité est le critère à l'aune duquel ces représentations sont évaluées : la question de savoir *comment* l'échantillon est produit est secondaire par rapport à sa correspondance aux normes. En d'autres termes, on peut dire que c'est la *justification* de la validité de la représentation qui permet d'assurer une continuité épistémologique et d'inscrire cette représentation dans les normes de la discipline. Dans l'ensemble, on peut considérer qu'il s'agit d'une épistémologie confirmationniste où un résultat est renforcé s'il concorde avec des résultats antérieurs.

De plus, le rapport à la norme méthodologique n'est pas absolument strict : les imperfections des différents modes de recueil sont connues et acceptées. Il n'est donc pas attendu d'un mode de recueil qu'il supprime toute forme de biais, même si une diminution de ceux-ci est bien sûr considérée comme une amélioration. L'adéquation à la norme n'étant jamais parfaite, la solution est de documenter les écarts de manière à ce qu'ils soient connus et pris en compte dans les analyses ultérieures.

En comparant le web à d'autres modes de recueil, les disciplines à terrain réalisent leur « conversion numérique » (Doueihi 2008) de manière conservatrice, en préservant leurs normes et leurs modalités de représentation du réel. L'avènement des *big data* n'a, à notre connaissance, pas suscité une appropriation massive des traces numériques par les disciplines à terrain, pour lesquelles la non-représentativité des traces constitue un obstacle intolérable. L'attachement aux normes est d'ailleurs vu par certains sociologues comme un danger de leur discipline qui se verrait « chassée de sa juridiction » par les éditeurs de plateforme en ligne. Il y a 10 ans, Mike Savage et Roger Burrows (2007) anticipaient déjà le risque d'un tel repli sur l'orthodoxie méthodologique :

> We can emphasize our superior reflexivity, theoretical sophistication, or critical edge. Fair enough – up to a point. Yet the danger is that this response involves taking refuge in the reassurance of our own internal world, our own assumed abilities to be more 'sophisticated', and thereby we chose to ignore the huge swathes of 'social data' that now proliferate.

En refusant de s'approprier les « social data » ou traces numériques, les disciplines à terrain réalisent leur mutation numérique avec une certaine aisance technique, mais sans révolution épistémologique. Comme nous allons maintenant le voir, cette mutation se fait de manière toute autre dans les disciplines à corpus, dont le médium a été réinventé par le numérique, et conduit à repenser non seulement les outils, mais surtout l'accès au réel à travers de nouveaux observables et de nouvelles méthodes pour les aborder.

4. Les défis de la révolution technique du texte

Les défis apportés par le numérique sont en effet tout autres pour les disciplines à corpus. Outre la révolution épistémologique qu'il introduit, le numérique constitue également un obstacle technique plus significatif dans ces disciplines, du fait du rapport qu'elles entretiennent avec la technique. Cette différence de rapport n'est pas sans rappeler celles des « deux cultures » de C.P. Snow (1959), celles des humanités d'un côté, et celle des sciences et techniques de l'autre, quoique la ligne de partage entre disciplines à corpus et à terrain ne coïncide pas tout à fait avec cette distinction. Comme on l'a vu, les disciplines à terrain sont plutôt d'obédience bifurcationniste et ne peuvent donc pas être assimilées aux sciences et techniques. Cette ligne de partage renvoie cependant à des rapports différents à la technique et au numérique au sein des sciences de la culture. Les disciplines à terrain se sont largement dotées d'outils produisant des résultats de nature quantitative, et ont réalisé progressivement leur « conversion numérique » d'une manière comparable aux sciences de la nature. Pour ces sciences de la culture déjà mathématisées, le numérique apparaît comme un moyen de démultiplier les possibilités, d'automatiser les tâches répétitives, de fabriquer de nouveaux observables, bref, un domaine à conquérir pour étendre le champ des possibilités de production de connaissances. Il n'y a pas de révolution épistémologique due à l'instrumentation technique de ces sciences de la culture quantitatives.

À l'inverse, les disciplines à corpus voient en effet leurs objets fondamentaux transformés par le numérique : avec la numérisation des archives, le document comme forme est arraché à son support analogique pour entrer dans un format numérique qui le rend toujours déjà manipulable et calculable, tandis que son support matériel (le disque dur) est séparé de son support de consultation (l'écran) (Bachimont 2010b). Le chercheur n'accède plus au document lui-même, mais à sa représentation, qui doit, pour bien jouer son rôle, présenter fidèlement les caractéristiques déterminantes de son objet. De même que l'écrit ne peut être conçu comme une pure transcription de la parole, le document numérique n'est pas une reproduction du livre ou de l'archive.

Comme l'a bien montré Christian Vandendorpe (1999), le texte n'est pas un pur contenu : il possède une sémiotique qui a connu ses propres mutations, et dont la forme actuelle constitue une forme de sédimentation ou de palimpseste. Cette sémiotique intègre des normes orthographiques, typographiques, ainsi que des conventions de renvois, d'index, de disposition en lignes, en paragraphes, en pages numérotées : tout ce qui peut paraître donné dans le livre imprimé est en réalité issu d'un lent détachement de l'oralité puis de l'écriture manuscrite, qui n'est pas sans impact sur le contenu lui-même. L'avènement d'une forme graphique permet des jeux de disposition (colonnes, encadrés, etc.), de mise en forme (graisse, italiques, typographie, etc.). Il introduit les nouvelles formes d'écritures possibles chères à Jack Goody (1979) : la liste, le tableau, la formule, auxquels Bruno Bachimont (2010b) propose d'adjoindre le schéma. Tous ces changements jouent un rôle dans

l'expérience perceptive de la lecture, sa compréhension, sa mémorisation, et même son appropriation : tout livre imprimé permet de souligner un passage, mais la taille des marges influe sur la propension et la facilité à inscrire des commentaires et notes de lecture. De la même manière, l'avènement du numérique (et tout particulièrement de l'hypertexte qui permet de renvoyer et naviguer dans le texte) transforme le format (ou média) du texte, avec des conséquences sur son contenu :

- le fil du texte est délinéarisé en séquences, en zones, en fragments ;
- sa clôture et son autonomie sont remis en question ;
- sa forme d'affichage est variable en fonction du dispositif d'affichage (ordinateur, écran, résolution, etc.) ;
- le texte peut proposer ou exiger des interventions du lecteur, et devenir interactif ;
- la forme textuelle n'est plus forcément dominante et à l'illustration du livre imprimé vient s'ajouter le son, la vidéo, le programme, etc. ;
- de nouvelles formes narratives non textuelles ou partiellement textuelles comme le jeu vidéo deviennent possibles.

Ces quelques exemples de mutations du texte avec le numérique, antérieures de quelques décennies à l'avènement des *big data*, se matérialisent pour le chercheur de deux façons :

- la numérisation de corpus transforme le texte imprimé en texte numérique, posant la nécessité de déterminer ce qui doit être préservé pour que l'on puisse dire que le texte numérique est une représentation viable du texte original ;
- de nouvelles formes textuelles se présentent au chercheur, et remettent également en cause les normes et les modalités du texte en général.

L'ontologie de la source se voit donc réinventée pour une population de chercheurs marqués par leur esprit de finesse, leur attention à la nuance et au détail, et pour lesquels de tels changements ne sont donc pas sans conséquences. L'avènement du numérique pour les humanités est un lieu spécifique du rapport à la technique, un lieu où la technique ne joue pas seulement le rôle de détermination et condition de possibilité matérielle de l'activité, mais se retrouve au cœur des modalités de configuration de l'activité elle-même.

Face à cet enjeu majeur, les acteurs des disciplines à corpus sont nombreux à envisager davantage la technique et le numérique comme un risque, un problème, ou tout du moins un sujet à discuter, que comme un espace d'outillage et d'expérimentation. C'est dans ces contextes que l'on trouve le plus de postures résolument technophobes. Certains de ces acteurs incarnent assez emblématiquement la culture humaniste, non technique, de C. P. Snow. Pour le dire de façon prosaïque, toute personne qui a assisté à une conversation entre une personne de formation purement littéraire et un ingénieur en informatique a conscience d'une certaine forme d'incommunicabilité entre le domaine des lettres et celui des techniques. Très concrètement, bon nombre de chercheurs éminents de ces domaines sont

désorientés par les outils basiques du numérique, ceux qui ne sont pas propres à l'activité de recherche : faire une recherche sur Internet, envoyer un mail, utiliser un outil de traitement de texte, ne vont pas forcément de soi, ont pu être adopté douloureusement. Cette difficulté a sans doute des explications : formation, tempérament, appréhension, etc., qu'il ne nous appartient pas de formuler ici. Cependant, on notera qu'elle peut se doubler d'une défiance, de formes de technophobie selon lesquelles la mobilisation de la technique se fait d'un façon toujours déjà problématique, dans une tradition déjà ancienne que jalonnent des auteurs comme Heidegger, Anders, Illich, ou Ellul en France. Ces réserves théoriques sur le numérique, doublées d'un manque de compétences pratiques à manipuler l'objet, sont emblématiques de la culture littéraire à l'œuvre dans les *digital humanities*, et de ses difficultés à se penser autrement que comme un projet de réflexion sur, et non avec, le numérique. Les hommes et femmes de lettres animés par ce projet se voient ainsi pauvrement armés face à l'enjeu ontologique et épistémologique majeur pour leur domaine de repenser ce qui supporte leur accès au réel, à savoir l'archive et le document, à l'aune du numérique. Il existe cependant plusieurs manières de faciliter cette transition.

Une façon de préparer cette révolution de l'accès au réel est un travail de modernisation des technologies d'organisation des savoirs, dont on a vu que les *big data* elles-mêmes font partie. À ce titre, les *digital humanities* ne sont pas toujours présentées comme un grand projet théorique de conversion numérique des humanités, mais aussi comme l'ambition plus pratique de rénover les outils de conservation, de traitement et de communications des savoirs. De ce point de vue, les exemples abondent : en France, des plateformes comme OpenEdition, Huma-Num, Cairn.info, Gallica, sont autant de dispositifs de publication, de diffusion, de partage et de consultation de connaissances à destination des chercheurs. À l'exception de certains aspects d'Huma-Num, il ne s'agit pas cependant d'outil de production de connaissances, ce qui se traduit également par la présence plus dominante d'autres types d'acteurs que des chercheurs. Les nouveaux modes de collaboration et de diffusion que ces outils induisent représentent des évolutions importantes du point de vue de la sociologie des sciences, en termes d'organisation et de fonctionnement des pratiques de recherche, mais sont de faible conséquence d'un point de vue épistémologique. Toujours à l'exception de quelques individus singuliers qui ont su faire la synthèse entre les « deux cultures » de C. P. Snow, les figures de ces plateformes sont davantage des ingénieurs (de recherche ou non), des informaticiens, des professionnels de l'information-documentation ou encore de la communication (voire du marketing). Leurs problématiques ne relèvent pas tant de la production de connaissances, que de leur préservation, de leur mise à disposition, de leur circulation.

À ce titre, un travail quantitatif mené en 2015 (Schmitt 2015) a mis en évidence la prégnance des problématiques de l'archive et de la bibliothèque dans les travaux de recherche effectivement publiés et revendiquant l'étiquette de *digital humanities* (voir **Figure** 2 et **Figure 3**). Il révèle que même en contexte académique, les *digital humanities* ne sont pas vécues comme une révolution dans la

production de connaissances en sciences sociales, et ne correspondent pas aux sciences computationnelles de la culture que nous avons conceptualisées précédemment : elles relèvent, encore une fois, des infrastructures de la recherche plus que de sa pratique. Le décalage entre les ambitions et la réalité des pratiques académiques en humanités numériques est d'ailleurs particulièrement marqué dans les *digital humanities* francophones, fortement resserrées sur l'examen des transformations du monde de l'édition et de l'information-documentation. À l'inverse, les *digital humanities* anglophones s'intéressent à des techniques computationnelles nativement numériques comme l'ingénierie des connaissances, le traitement du langage, les systèmes d'information géographiques, etc., qui permettent de mettre au point de nouveaux outils de production de connaissances. Le versant anglophone apparaît ainsi comme plus avancé dans sa réflexion sur l'outillage des humanités, avec, semble-t-il, le parti pris d'expérimenter avec et sur les outils computationnels existants, fût-ce de manière un peu déstructurée et opportuniste, plutôt que de les problématiser *a priori* sans les avoir effectivement utilisés.

Une autre façon d'intégrer les « nouveaux » outils numérique dans la production de connaissance des disciplines à corpus est de s'appuyer sur une division du travail qui fasse collaborer des chercheurs en sciences de la culture et des informaticiens. Les *digital humanities* se donnent en effet comme ambition récurrente de promouvoir l'interdisciplinarité entre sciences humaines et sociales, mais aussi et peut-être surtout avec l'informatique : l'intérêt pour les outils numériques unirait les sciences humaines et sociales avec une problématique commune, et conduirait à une rencontre avec l'informatique. Du point de vue de notre recherche de continuité épistémique, cette interdisciplinarité est apparue comme une nécessité pour dépasser la division du travail entre chercheurs en sciences de la culture et profils techniques, de manière à ce qu'une épistémologie commune puisse régir l'activité de ces différents acteurs. Sur le premier point, l'interdisciplinarité au sein des sciences de la culture, Marin Dacos et Pierre Mounier écrivent en effet dans leur état des lieux sur les *digital humanities* (2014) :

> Au plus haut niveau de généralité, on pourrait dire que les humanités numériques désignent un dialogue interdisciplinaire sur la dimension numérique des recherches en sciences humaines et sociales, au niveau des outils, des méthodes, des objets d'études et des modes de communication. »

La réflexion sur les objets et les outils induite par l'avènement du numérique pourrait ainsi être menée en commun par les différentes disciplines mobilisées dans les sciences de la culture, unies par un même enjeu de penser son rapport à la technique et au numérique. Néanmoins, la possibilité d'une réflexion commune est remise en cause par la tension que nous avons mise en évidence entre les disciplines à corpus et à terrain, et les rapports respectifs qu'elles entretiennent à la technique : il y a entre un historien et un économiste des différences de culture épistémique que les meilleures intentions du monde peuvent difficilement espérer combler. Ce premier point nous semble donc peu convaincant, et à tout le moins limité par les deux cultures des sciences humaines et sociales.

Sur le second point (interdisciplinarité entre sciences de la culture et informatique), il apparaît que les conditions requises pour donner naissance à une interdisciplinarité se présentent rarement. En effet, le fait de passer commande pour une application web, d'acheter un logiciel, de le faire développer par un ingénieur de recherche, ne crée pas une situation d'interdisciplinarité mais un rapport hiérarchique entre deux types d'acteurs, les fournisseurs et les utilisateurs. Bien que la recherche en design de ces dernières années souligne régulièrement l'importance de la collaboration et des principes de cocréation pour concevoir avec succès des outils numériques, cette configuration collaborative ne correspond pas à un dialogue scientifique entre disciplines. Pour qu'un tel dialogue ait lieu, il faut en effet que l'informatique soit convoquée en tant que discipline scientifique. Or, ces « sciences de l'artéfacture » (Bachimont 1996) ou « science des artefacts interactifs » (Lassègue 1996) désignent un domaine vaste qui va de la recherche théorique sur l'information à des champs d'application multiples, dont l'informatique en tant que secteur d'activité économique. Faire développer une application web ne mobilise pas des chercheurs en informatique mais des ingénieurs et développeurs qui utilisent des outils et méthodes originellement fondés sur les principes mis en évidence par la recherche en informatique : entre ces deux activités, il y a une distance qu'on peut comparer à celle qui existe entre l'établissement des principes de la thermodynamique classique au XIXe siècle, et le fait d'installer un radiateur dans une université. Dans les *digital humanities*, l'informatique n'est généralement pas convoquée comme discipline scientifique, mais instrumentée comme une technique au service de la production de savoir : rares d'ailleurs sont les équipes qui associent des chercheurs venant des deux champs disciplinaires. Or, comme le souligne Aurélien Bénel (2013) dans un article sur l'interdisciplinarité dans les *digital humanities*,

> il n'y a pas d'interdisciplinarité sans engagement des chercheurs des deux disciplines (ou plus), pas d'interdisciplinarité réussie sans enrichissement mutuel.

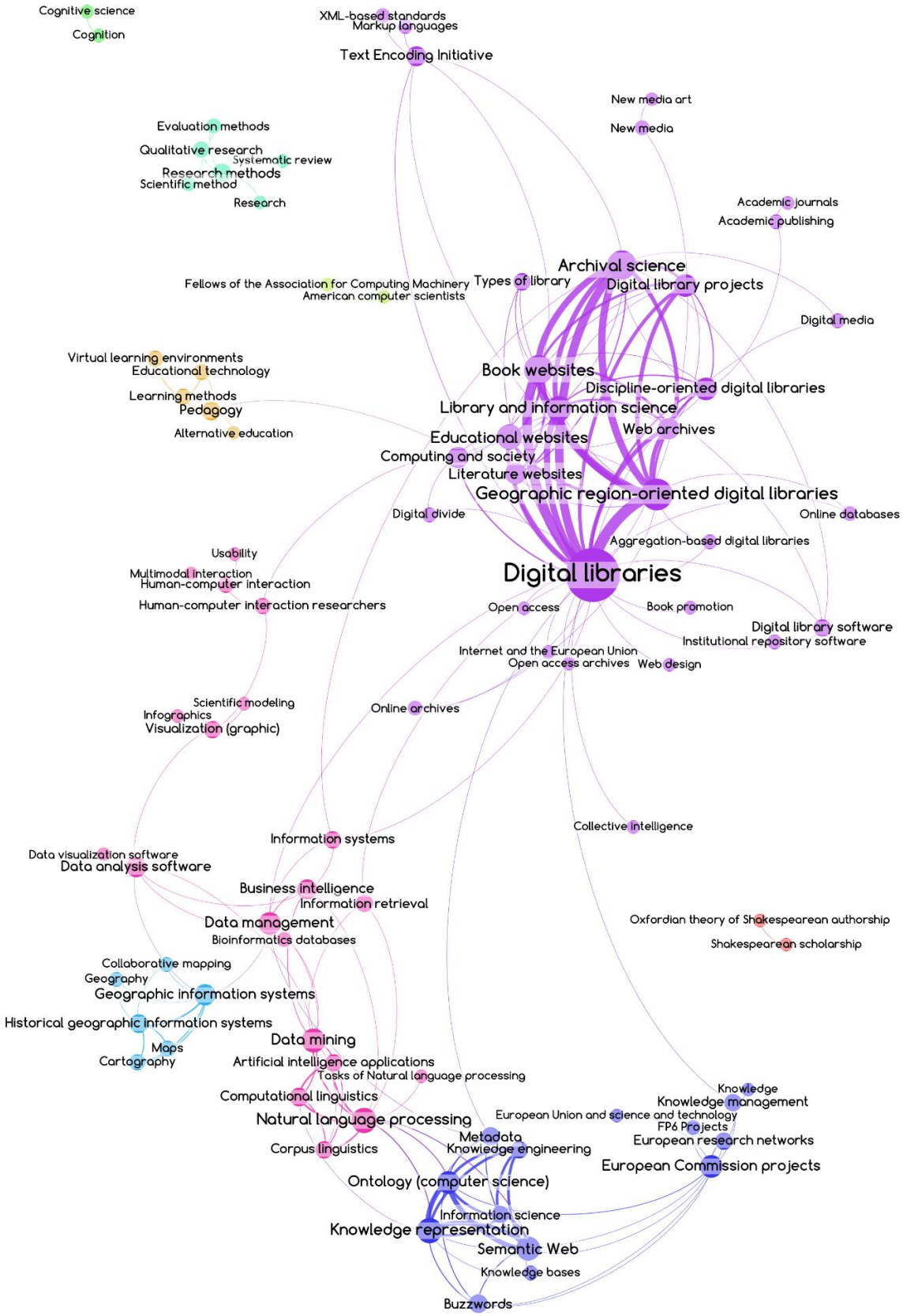

Figure 2. *Cartographie thématique des* digital humanities *anglophones, d'après Schmitt (2015).*

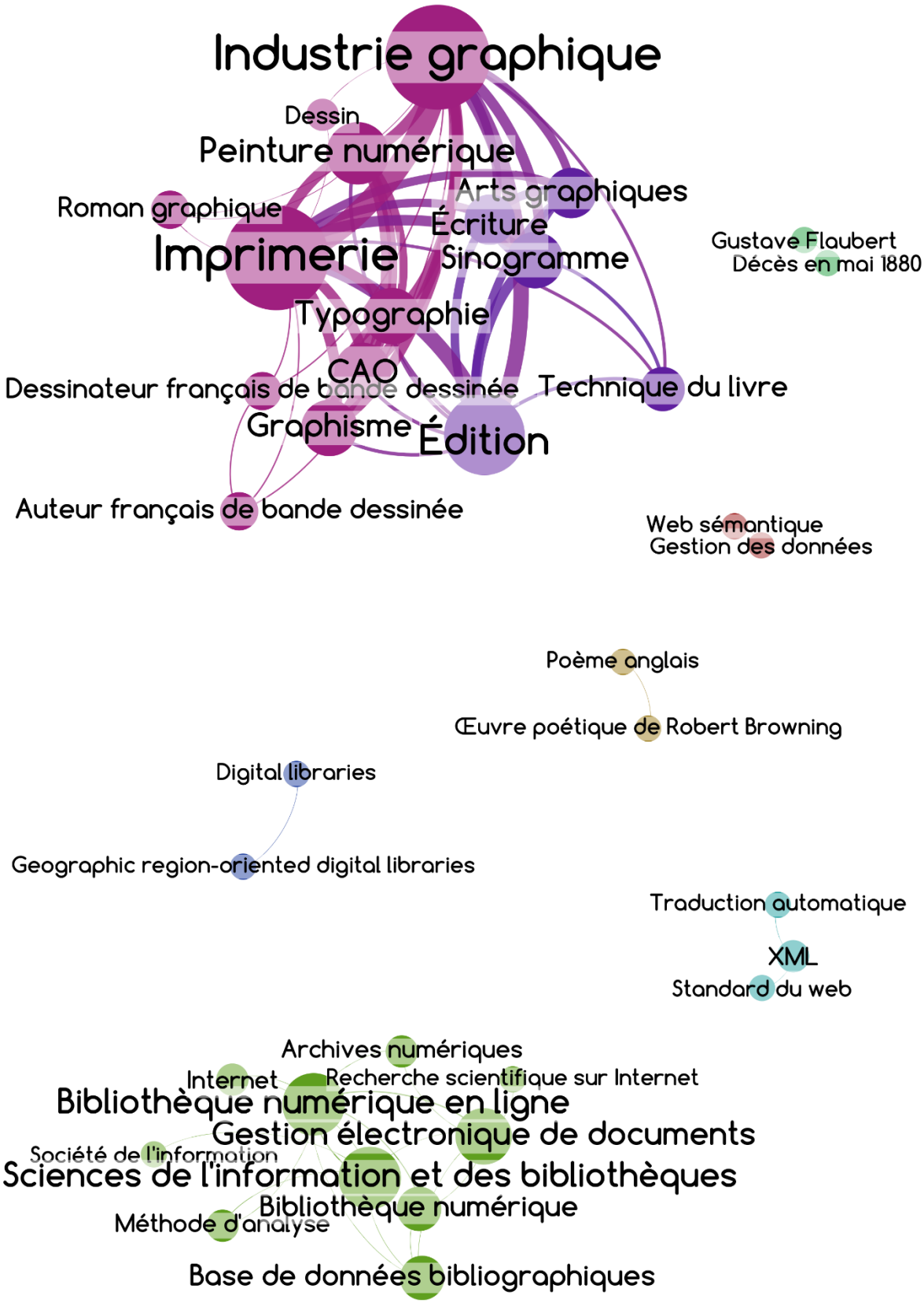

Figure 3. *Cartographie thématique des* digital humanities *francophones, d'après Schmitt (2015).*

La défiance envers les technologies numériques et les difficultés à construire une configuration interdisciplinaire entre l'informatique et les sciences de la culture expliquent largement le caractère encore embryonnaire du développement des *digital humanities*. Dans les disciplines à corpus, l'irruption du numérique fait intervenir une révolution technique doublée d'une révolution épistémologique qui renouvelle à la fois le mode d'accès au réel et les outils pour exploiter les observables, soulevant des défis largement supérieurs à ceux que rencontrent les disciplines à terrain.

Ce constat présente cependant une exception notable, celle de la linguistique. Si le mode d'accès au réel de la linguistique est résolument le corpus, son approche est plutôt naturaliste dans l'ensemble et plus compatible avec la culture des sciences et techniques. Il existe du moins une branche de la linguistique fortement naturalisée et hybridée avec les technologies numériques. À l'intersection des cultures textuelles et quantitativistes, des champs comme la linguistique computationnelle, mais aussi la stylométrie, le *distant reading* en histoire, la textométrie, se proposent de rendre calculable le texte envisagé comme donnée. La particularité de la linguistique computationnelle est qu'elle est, à l'instar de l'informatique en général, à la fois un domaine de recherche en soi et une science appliquée, voire une technologie, mobilisable par d'autres disciplines, dont l'analyse littéraire, l'histoire, mais aussi la sociologie. Le cadre théorique de l'analyse computationnelle est celui de la discipline cible (et non la linguistique) qui régit à la fois la logique de constitution des observables, la méthode et les outils : il y a ici une continuité épistémique à la fois instrumentée par le numérique et compatible avec les sciences de la culture, dont le défaut est néanmoins qu'elle repose sur l'existence de corpus homogènes préexistants, dont elle ne prend pas en charge la cohérence, et ne peut donc s'appliquer telle quelle aux traces numériques.

Elle permet ainsi de revisiter des unités historiquement et culturellement attestées, comme l'a fait Roberto Busa avec le *Corpus Thomisticus*. Un autre exemple emblématique de cette redécouverte d'objets familiers dans le champ de l'analyse littéraire est l'expérience de textométrie appliquée à un texte de Maupassant rapportée par François Rastier (2001) :

> alors qu'il m'avait fallu dix années pour comprendre l'importance du nombre *dix* dans la nouvelle de Maupassant intitulée *Toine* (l'auteur, 1989, liv.II, chap.V), le test de l'écart réduit me l'a instantanément mis sous les yeux, et m'a même permis de tirer parti d'une occurrence à la première ligne, qui m'avait, je l'avoue, échappé, bien qu'elle eût renforcé mon propos.

Dans cet exemple, l'analyse computationnelle du texte de Maupassant permet de reconstituer presque instantanément une connaissance déjà établie ; on peut imaginer que la même démarche permettrait de produire des connaissances nouvelles sur un corpus moins familier. La validité de cette connaissance repose tout autant sur le résultat d'un calcul que sur la capacité de François Rastier à donner une signification linguistique et littéraire à ce résultat. La linguistique computationnelle permet ainsi de redécouvrir des corpus établis, et de reproblématiser les objets des sciences de la culture à travers leur numérisation et leur caractère nouvellement manipulable.

Cette approche textométrique peut être appliquée dans des disciplines à terrain comme la sociologie : la statistique textuelle pour les sciences sociales, ou analyse statistique de données textuelles (ASDT), donne au corpus le rôle d'un intermédiaire pour observer des faits sociaux (et non textuels). Le corpus devient un médium pour accéder au terrain, faisant la synthèse des deux familles de disciplines que nous avons distinguées plus haut.

Cette démarche est notamment utilisée pour analyser les retranscriptions d'entretiens qualitatifs, et les réponses aux questions ouvertes des sondages d'opinion. Des logiciels comme Alceste (Reinert 1990) ou Prospéro (Chateauraynaud 2003) mobilisent le cadre conceptuel de la sociologie quantitative pour conceptualiser des observables linguistiques et des opérations de transformation computationnelle de ces observables compatibles avec les normes de la discipline. Max Reinert s'inspire ainsi de l'école française de statistique développée par Jean-Paul Benzécri et de la linguistique distributionnelle de Zellig Harris. Il dispose ainsi d'une méthode d'articulation entre linguistique, statistique, et sociologie, qui préexiste au logiciel. Les données linguistiques sont généralement issues des réponses à des questions ouvertes : elles héritent donc d'une épistémologie quantitativiste fondée sur la notion d'échantillon représentatif qui les rend compatibles avec les exigences de la sociologie quantitative. Pour les exploiter, Reinert et Chateauraynaud développent une théorie de l'instrument (le logiciel) qui hérite de leur culture épistémique. L'instrument permet de découper des unités linguistiques et de calculer des fréquences linguistiques qui ont un statut d'observations quantitatives, et qui peuvent dès lors être analysées et interprétées suivant la logique de la sociologie quantitative. Cette analyse est possible d'une part parce que les données correspondent aux normes de la discipline, et d'autre part parce qu'il y a une continuité épistémique entre ces données, le logiciel, et la méthode qui inclut l'utilisation du logiciel et l'interprétation des résultats.

La présence de continuité épistémique atténue également la problématique de la maîtrise de la théorie de l'instrument par les utilisateurs du logiciel. Cette problématique est relativement classique en philosophie des sciences (Adam 2004; Chalmers 2003; Franklin 1994; Schindler 2013), avec en synthèse plusieurs positions allant de la nécessité de connaître les normes théoriques qui ont préludé à la mise au point de l'instrument pour pouvoir l'utiliser, à la possibilité de l'utiliser comme une boîte noire pourvu qu'on sache interpréter les résultats qu'il produit. Soulignons par ailleurs que le débat épistémologique porte surtout sur le risque de circularité de l'utilisation pour prouver une théorie d'instruments construits sur la base de cette même théorie. Dans notre perspective, la question de déterminer s'il faut par exemple savoir comment fonctionne un microscope pour en utiliser un n'appelle naturellement pas une réponse binaire. Affirmer que seul un microscopiste pourrait utiliser un microscope car seul le concepteur aurait cette maîtrise complète de la théorie de l'instrument, serait intolérable en termes d'utilité de l'objet. À l'inverse, l'idée que n'importe qui peut voir quelque chose dans un microscope est contredite par l'expérience, comme le confirmera n'importe quel

biologiste en herbe ayant manipulé un microscope non réglé sur son échantillon. La question de savoir si l'utilisateur de l'instrument doit en maîtriser la théorie dépend donc de l'objet, du contexte disciplinaire, de la maturité technologique de l'instrument, etc.

Dans le cas des logiciels de statistique textuelle pour les sciences sociales, l'instrument est fourni avec une méthode d'utilisation (présentée notamment dans une publication scientifique) qui découle de la théorie de l'instrument sans forcément l'expliciter ; cette méthode fait l'objet de plusieurs publications présentant l'approche générale, ou un cas d'utilisation du logiciel. La continuité disciplinaire entre le concepteur de l'instrument et ses utilisateurs donne aux éléments proposés une certaine intelligibilité intuitive dès lors que l'on a pris en main les opérations techniques à effectuer.

L'analyse statistique de données textuelles combine ainsi le cadre théorique et méthodologique des sciences de la culture avec la capacité à mobiliser des données numériques et des outils de traitement computationnel. Elle constitue donc un candidat intéressant pour réaliser le programme des sciences computationnelles de la culture. Elle présente en effet une continuité entre la donnée, l'outil, la méthode et le cadre théorique qui correspond aux critères que nous avions établis pour cette discipline putative. Sa lacune du point de vue des sciences computationnelles de la culture provient de la nature des observables qu'elle manipule : l'analyse statistique de données textuelles reste soumise, en termes de mode d'accès au réel, aux normes des différentes disciplines qui la mobilise. L'analyse doit en effet se faire sur un échantillon représentatif (en sociologie quantitative) ou un corpus homogène (dans les disciplines à corpus). Elle ne peut donc pas être mobilisée telle quelle pour l'exploitation des traces numériques, pour lesquelles les disciplines à corpus et à terrain n'ont pas de proposition de constitution d'un ensemble normé.

Cette absence de proposition ne veut pas dire qu'aucun travail de recherche n'a été réalisé pour mobiliser les traces numériques dans les sciences de la culture. Il nous montre en revanche qu'il n'y a pas de candidat existant proposant à la fois une épistémologie de la trace numérique et une cohérence avec les outils, les méthodes et le cadre théorique : pour notre paradigme méthodologique, il nous faudra combiner plusieurs candidats ou composantes. Pour cela, nous examinons maintenant les projets de recherche proposant une épistémologie de la trace numérique non pas générale, comme nous l'avons formulée au chapitre III, mais appliquée à un objet spécifique, empiriquement constitué.

Cet objet sera, à nouveau, l'exemple du web. À travers cet exemple on verra que les traces numériques sont bien un fait humain qui intéresse ces disciplines. Néanmoins, ces traces sont avant tout envisagées comme un objet en soi, et non comme un intermédiaire représentationnel pour accéder au fait humanité dans sa globalité. Ce n'est qu'ultérieurement, dans l'histoire des « sciences du web », que cet objet désormais mieux connu cesse d'être simplement un objet « virtuel » et commence à être considéré comme une manière d'accéder au « réel ».

5. Du web comme objet au web comme corpus

Les traces du web n'ont pas toujours été considérées comme des sources pour l'analyse de phénomènes socioculturels en général, qu'ils soient numériques ou « analogiques ». Le web peut s'entendre comme un objet d'étude en tant que tel, qui bénéficie de méthodes d'archivage et d'une historiographie propre. Il est en lui-même un objet foisonnant, fait de contenus, de langages, de protocoles, et servant à s'étudier lui-même grâce à sa propre positivité matérielle. Avec la propension du monde anglo-saxon à nommer de nouveaux champs de recherche autour d'un objet, l'étude du web s'inscrit dans une discipline, la science du web *(web science)* :

> Over the past few years, there has been a growing recognition that the ecosystem that is the Web needs to be treated as an important and coherent area of study—this is Web science. (Shadbolt et al. 2013)

Le web est un dispositif technique, fait de choix qui peuvent être étudiés ou modifiés : la *web science* s'envisage tout à la fois comme un domaine d'étude de son objet et comme un terrain d'amélioration et de modification de celui-ci, le web n'étant pas figé d'un point de vue technique :

> Physical science is an analytic discipline that aims to find laws that generate or explain observed phenomena; computer science is predominantly (though not exclusively) synthetic, in that formalisms and algorithms are created in order to support particular desired behaviour. Web science has to be a merging of these two paradigms; the Web needs to be studied and understood, and it needs to be engineered. (Berners-Lee et al. 2006)

Qu'il s'agisse de l'évolution de ses formats (révision majeure du HTML avec la version 5 en 2014[21]) ou d'innovations majeures comme les Linked Data (O'Hara et Hall 2012), le web est un objet dynamique, qui s'envisage d'un point de vue synchronique aussi bien que diachronique. En termes d'approches, de nombreuses représentations du web sont possibles, mobilisant un kaléidoscope de disciplines. La question de la structure mathématique ou logique du web a mobilisé un certain nombre de chercheurs dans différents domaines, avec plus ou moins de sérieux. Les travaux sur la topologie du web d'Albert-Lászlo Barabási (1999; 2013) et de ses collègues ont eu beaucoup d'influence, en particulier sur les méthodes et outils d'analyse de graphe appliqués au web ; à l'opposé, les tentatives pour définir la forme du web de manière intuitive, comme un nœud papillon, une théière ou un chou-fleur (Metaxas 2012) ont moins d'échos universitaires et médiatiques. Au-delà de l'anecdote, ce genre de travaux traduit un foisonnement d'axes de recherche qui sont loin de former un tout cohérent et dans lesquels il est difficile de s'orienter. À cela s'ajoute le fait que le web, comme tout objet technique, est le support d'usages prévus ou imprévus, et que l'on peut aussi bien s'intéresser à ce qu'il représente culturellement et socialement qu'aux raffinements de sa matérialité technique.

[21] https://www.w3.org/TR/html5/ (consulté le 16 novembre 2017)

D'un point de vue socioculturel, l'originalité du web par rapport à d'autres environnements techniques est, comme on l'a vu, qu'il garde la trace de presque tout ce qui s'y produit. Le web est pour ainsi dire constitué de ce qu'il laisse derrière lui et qui, contrairement aux ruines antiques, semblent ne jamais s'altérer. Pour le chercheur en sciences de la culture, cela veut dire qu'il n'a pas besoin d'assister aux événements qui s'y produisent, ou de les reconstituer à travers des témoignages, pour en garder la trace. Le web est sa propre trace, sa propre archive, ce qui ne veut pas dire que son étude soit dépourvue d'enjeux historiographique et philologique. Comme l'ont souligné notamment Alexandre Monnin (2012) et Niels Brügger (2012), l'enregistrement par défaut n'est pas synonyme de stabilité : la page d'accueil du Monde chez l'un, ou de TF1 chez l'autre, sont des exemples de contenus qui changent d'une heure à l'autre, s'affichent différemment en fonction du matériel de l'internaute, et peuvent par ailleurs être personnalisés pour l'internaute lui-même. Le support numérique ne se dégrade pas comme le support papier, avec une perte graduelle de l'information qui y figure ; en revanche, les formats numériques évoluent et peuvent devenir rapidement obsolètes.

L'archivage exhaustif du web n'étant pas envisageable, il est nécessaire de se poser la question des unités à archiver et de proposer une ontologie au moins instrumentale du web, c'est-à-dire de définir quels sont ses objets : page, site, ressource, lien hypertexte, document, contenu, etc. Encore moins que dans la philologie médiévale n'y a-t-il de document « original » que l'on pourrait reconstituer à partir de ses variations. La page web n'est pas un document que l'on peut établir ; l'impermanence de sa forme et de son contenu lui est constitutive. Il est discutable que l'on puisse lui attribuer une existence essentielle au-delà de ses différentes représentations. C'est en effet un objet graphique, textuel, computationnel, technique, qui doit être réduit à l'une de ces dimensions pour pouvoir être archivé et étudié dans un format spécifique, par exemple :

- des captures d'écran pour l'étudier graphiquement ;
- une indexation plein texte pour l'étudier textuellement ;
- une copie du code source de la page ou de la base de données pour l'étudier techniquement ;
- etc.

Tout archivage est donc nécessairement partiel et déterminé par des décisions méthodologiques, elles-mêmes formulées au regard d'un certain nombre d'usages prévisibles. Ce que nous avons évoqué à l'échelle d'un objet comme la page se retrouve à l'échelle du web en général. Il n'est pas question d'archiver l'intégralité du web mais d'élaborer des stratégies de collecte qui s'effectuent selon une logique cohérente. Les collectes ciblées, comme celle de Guilhem Fouetillou (2007) à propos du traité constitutionnel européen, visent une « localité du web » qui s'articule en termes de thématiques instanciées par des mots-clés de recherche et des sites web de référence. À partir de ces sites référents sélectionnés par des experts de la thématique suivant des critères éditoriaux, la collecte se fait par rebond, en suivant les liens hypertextes présents dans les pages de ces sites, de manière à délimiter un espace de pages autour de ces sites référents, dont la distance se mesure en nombre de liens d'écart

aux sites référents. C'est une collecte synchronique, qui constitue un corpus à un moment donné, et ne s'appuie pas sur la temporalité du web mais sur sa spatialité réticulaire et les applications qu'elle propose en matière d'analyse de graphe. À l'inverse, le dépôt légal du web proposé par l'Ina collecte des pages sur une liste prédéterminée de quelques milliers de sites, mais à intervalles temporels réguliers, avec une inscription dans la durée (Mussou 2012). Internet Archive présente une stratégie encore différente, avec une ambition initiale plutôt de l'ordre de la bibliothèque ou du fonds documentaire, fondée sur la curation ; aujourd'hui cette ambition a évolué notamment du fait de sa dimension collaborative, par laquelle tout visiteur d'Internet Archive peut mettre sous surveillance un site web qui sera désormais archivé à intervalles réguliers. Les robots indexeurs des moteurs de recherche reposent sur des stratégies de collecte encore distinctes, qui agrègent les précédentes avec d'autres techniques spécifiques.

L'archivage du web dans sa spatio-temporalité est une problématique qui mobilise de multiples cultures disciplinaires au sein des sciences de la culture, comme l'histoire, la philologie, l'étude des médias. Il se fait généralement au service d'un projet spécifique, dans l'idée d'étudier un objet particulier à travers une « localité du web ». Historiquement, les premières initiatives comme Internet Archive dans les années 1990 sont influencées à la fois par des mondes connus comme celui de la bibliothèque, mais aussi par un imaginaire propre au web, dans lequel il n'est pas seulement une spatio-temporalité sans forme dans laquelle on peut puiser, mais un objet conceptualisé et essentialisé comme un territoire. Or, comme le souligne Pierre Musso (2010) :

> Un territoire n'est pas simplement un espace, mais la représentation collective d'un espace-temps, un lieu d'histoire et de projets "enracinés", ancrés dans un espace.

Cet imaginaire du web comme objet-territoire porte sa propre fécondité et a imprégné à la fois un certain nombre de domaines de recherche, mais aussi la façon dont le web a pu être étudié jusqu'à un certain point. Dans les *cybercultural studies* (citées par Richard Rogers, 2010) et les études du cyberespace (Graham 2013), le web est un objet à part entière, envisagé dans sa globalité, dont la matérialité est constituée des traces qui servent à l'étudier. Dans ces deux domaines de recherche, caractéristiques de la fin des années 1990, on retrouve chaque fois un imaginaire de la spatialité qui postule que le web est un espace physiquement ou virtuellement séparé, déconnecté du « vrai » espace géographique (Andrade 2010). La métaphore du monde parallèle apparaît comme une façon de s'approprier culturellement et linguistiquement une technologie nouvelle. Rey (2012), cité par Graham (2013), soutient que le cyberespace était une fiction nécessaire pour surpasser la dissonance cognitive due à l'idée de communications instantanées bien qu'à distance, rendues possibles par un espace intermédiaire. Comme dans l'imaginaire des nanotechnologies (Loeve 2009), les discours de science-fiction sont un élément constitutif de la conceptualisation de l'objet.

Le genre littéraire *cyberpunk* renvoie ainsi à un certain nombre d'œuvres de fiction emblématiques de ce besoin d'appropriation de la technologie. L'expression « cyberespace » a été forgée en 1984 par l'auteur de science-fiction William Gibson dans son roman *Neuromancer* où il le décrit comme une

> hallucination consensuelle vécue quotidiennement en toute légalité par des dizaines de millions d'opérateurs, dans tous les pays, par des gosses auxquels on enseigne les concepts mathématiques… Une représentation graphique de données extraites des mémoires de tous les ordinateurs du système humain. Une complexité impensable. Des traits de lumière disposés dans le non-espace de l'esprit, des amas et des constellations de données. Comme les lumières de villes, dans le lointain.[22]

D'autres œuvres imaginent des espaces virtuels, comme l' « infosphère », la « datasphère » voire la « matrice », qui sont autant d'avatars d'un même imaginaire idéologiquement et politiquement chargé. En effet, derrière la « vue d'artiste » du cyberespace à la Gibson, on retrouve un univers utopiste et libertaire emblématiquement représenté par la « déclaration d'indépendance du cyberespace » du poète et militant John Perry Barlow (1996), qui présente le web comme un nouveau territoire sans maître, plus libre et plus juste, indépendant de la souveraineté des états, avec sa culture, son éthique, sa civilisation.

Cet imaginaire, artistique et politique, déborde du cadre des travaux universitaires portant sur le web, mais les a durablement influencés dans la représentation qu'ils se font de leur objet. La question du virtuel n'est pas tant une question épistémologique qu'ontologique : le virtuel caractérise l'objet lui-même. Or le cyberespace n'est pas tant un objet virtuel qu'une fiction. L'idée qu'il y aurait entre le web (ou tout autre dispositif permettant d'être « en ligne ») et le « vrai » monde, une séparation radicale, est aujourd'hui contestée, et pour le moins considérée comme obsolète. Nathan Jurgenson (2012) dénonce ainsi le *dualisme numérique* qui consiste à considérer le « virtuel » et le « réel » comme ontologiquement séparés, comme si les réseaux sociaux en ligne et les personnes qui y écrivent étaient deux réalités distinctes. Il oppose à ce dualisme l'idée d'une *réalité augmentée* par le numérique, en prenant l'exemple des mouvements de protestation comme les révolutions arabes qui ne se sont pas déroulées séparément sur Internet et dans les pays arabes, mais dans une réalité qui inclut ces deux espaces. Le passage d'un web plutôt documentaire à un web peuplé d'interactions sociales a vraisemblablement facilité ou accéléré le rejet du dualisme numérique porté par l'imaginaire du cyberespace, dans la mesure où il est difficile de maintenir une distinction ontologique entre les interactions sociales en ligne et les interactions sociales en général.

Cette évolution des représentations mentales du web trouve aujourd'hui sa traduction dans la manière d'étudier le web. S'il est toujours bien question d' « espaces » en ligne, c'est désormais dans un sens plus abstrait. Les espaces en ligne ne sont plus virtuels, et les échanges sociaux qui se déroulent dans

[22] « A consensual hallucination experienced daily by billions of legitimate operators, in every nation, by children being taught mathematical concepts... A graphic representation of data abstracted from the banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding. » Traduction de Jean Bonnefoy, éditions La Découverte, Paris, 1985

les communautés en ligne ne sont pas ontologiquement différents des échanges sociaux tels que la sociologie les a toujours connus ; connaître quelqu'un sur Internet est une sorte de façon de connaître quelqu'un en général. Historiquement, cette séparation se présente de façon apodictique, comme un postulat et non comme une explication : influencée par un imaginaire, rien ne la justifie plus d'un point de vue conceptuel. Le dépassement de cette séparation permet notamment d'envisager le web comme une représentation de la culture, et ouvre la voie à une « recherche numérique » *(digital research)* qui envisage le web comme la trace d'autre chose.

Il existe ainsi plusieurs projets de recherche inscrits dans les sciences de la culture et posant la question de l'exploitation du web comme trace du fait humain, dont notamment les *digital methods* et les *cultural analytics*. Ces différents champs ont en commun de renvoyer à des groupes de quelques dizaines de chercheurs, organisés dans une structure de recherche (respectivement la *Digital Methods Initiative* à Amsterdam et la *Software Studies Initiative* distribuée entre New York et La Jolla (Californie)). Ces structures, qui ont aujourd'hui une dizaine d'années, ont, à nos yeux, travaillé dans deux sens complémentaires : une conceptualisation du web et des réseaux sociaux dans la perspective des sciences de la culture, et l'expérimentation autour d'outils « maison » de collecte et d'analyse de traces numériques. Dans les deux cas, il ne s'agit pas vraiment de « faire école » et d'imposer des normes d'utilisation des outils : les logiciels sont mis à disposition tels quels, avec une contextualisation parfois laconique, sans qu'il faille s'inscrire dans la conceptualisation du web proposé par ces structures. En d'autres termes, bien que ces deux chantiers soient complémentaires, ils existent de manière relativement autonome. Il nous semble également qu'ils correspondent à deux étapes de maturité de ces structures : d'abord une phase ambitieuse de théorisation de l'objet, qui n'a, dans les deux cas, pas vraiment abouti, suivi d'une phase plus pragmatique d'expérimentation visant à « faire pour comprendre » plutôt qu'à « comprendre pour faire ».

Il y a une dizaine d'années, Richard Rogers (2009) a ainsi proposé, sur un mode plus prospectif que rétrospectif, des études sur Internet qui seraient des outils de diagnostic des changements culturels et sociaux, qu'ils se produisent en ligne ou hors ligne :

> I would like to put forward a new era in Internet research, which no longer concerns itself with the divide between the real and the virtual. It concerns a shift in the kinds of questions put to the study of the Internet. The Internet is employed as a site of research for far more than just online culture. The issue no longer is how much of society and culture is online, but rather how to diagnose cultural change and societal conditions with the Internet.
>
> [I]n studying the online, we make and ground findings about society and culture with the Internet (Rogers 2010)

Dans le même esprit, Lev Manovich (2007) inaugure le terme de *cultural analytics* pour désigner de nouvelles manières de connaître et comprendre les phénomènes culturels :

> We feel that the ground has been set to start thinking of culture as data (including media content and people's creative and social activities around this content) that can be mined and visualized. In other words, if data analysis, data mining, and visualization have been adopted by scientists, businesses, and government agencies as a new way to generate knowledge, let us apply the same approach to understanding culture.

Ces nouvelles voies de recherche se distinguent des *virtual methods* évoquées plus haut : Richard Rogers insiste sur le caractère nativement numérique des données et des méthodes correspondantes. Il n'y a pas de continuité méthodologique avec l'existant : l'objet suscite de nouvelles méthodes. Il n'est donc pas question de reconstituer une représentativité des données comme dans les *virtual methods* : il est illusoire d'évaluer ces nouvelles méthodes à l'aune de la normativité des méthodes classiques. Les normes méthodologiques telles que la recherche de représentativité statistique ne permettent pas de s'approprier les traces numériques du web. Les sources de données, mais aussi les outils et les concepts pour les appréhender, sont renouvelés. Aux catégories sociodémographiques habituelles des sciences sociales, Richard Rogers (2008) oppose les classes « post-démographiques » propres au support numérique comme les centres d'intérêt déclarés, les abonnements, les bibliothèques de médias, les contenus produits par les individus. Ces catégories, plus adaptées à l'analyse de données web, ne servent plus à redresser statistiquement les données comme le permettaient les caractéristiques sociodémographiques, mais à faire émerger de nouvelles typologies d'individus, et de manière productive, à leur proposer des contenus (qu'il s'agisse de recommandation musicale, de curation de contenu ou de ciblage publicitaire). Elles sont associées à une redistribution des acteurs qui ne sont plus seulement les chercheurs mais aussi les plateformes qui recueillent ces catégories, et la multitude d'utilisateurs qui les alimentent et les exploitent.

Un autre aspect du projet de recherche de Richard Rogers est l'intérêt porté au *matériau*. Dans sa conférence inaugurale de 2009, il propose comme mot d'ordre de *suivre le matériau* (« follow the medium »). Du fait des mutations permanentes des objets numériques, dont les conceptualisations se désagrègent rapidement, il suggère ainsi de ne pas faire partir les recherches des objets à étudier mais des observables dont on dispose sur le moment. Dans cette approche empiriste, c'est *a posteriori* que l'on détermine si ces observables sont des représentations du web lui-même (dans sa totalité ou ses parties) ou de phénomènes socio-culturels plus larges. La valeur épistémique de la donnée web est déterminée « sur pièce », après examen, et non *a priori* lorsqu'un certain genre de donnée est choisi par rapport à un objet de recherche prédéterminé.

Toutefois, cet empirisme méthodologique reste ancré dans la culture épistémique des sciences humaines et sociales : il reste critique par rapport aux sources, aux méthodes, aux outils. Le fait de partir du matériau plutôt que de l'objet ménage une place à l'innovation méthodologique, à l'ouverture des questions de recherche, à une approche exploratoire qui part de fondements empiriques, mais revient néanmoins à une évaluation critique. C'est également un empirisme qui se donne *a priori* la culture et la société comme horizon, et n'explore pas de données pour elles-mêmes,

mais dans l'idée de leur trouver une valeur épistémique relative à des faits socioculturels ; Richard Rogers résume sa démarche de façon provocatrice en proposant de voir la société dans les résultats de recherche de Google (« we look at Google results and see society, instead of Google », Richard Rogers, 2013). Le web n'y a pas le statut d'un objet à part, pouvant être étudié dans ses dimensions techniques et culturelles en tant que telles, mais un statut de source d'observation et d'étude du fait humain dans sa globalité.

Cet empirisme méthodologique, comme on l'a vu, reste cependant incompatible avec l'orthodoxie méthodologique des sciences de la culture : Richard Rogers ne propose pas de solution pour restaurer la continuité avec le paradigme de l'échantillonnage et de la représentativité, ni de théorie de l'archive à proprement parler. En termes de culture épistémique, il s'inscrit, notamment avec l'idée de « suivre le matériau », dans la continuité des *media studies* :

> following the medium is a particular form of medium-specific research. Medium specificity is not only how one subdivides disciplinary commitments in media studies according to the primary objects of study: film, radio, television, etc. It also refers to media's ontological distinctiveness, though the means by which the ontologies are built differ. (*Ibid.*)

Or un média est, par définition, un objet intermédiaire, qui relie un contenu et son lecteur (ou spectateur). Considérer le web comme un média permet en effet de l'inscrire dans la culture disciplinaire des *media studies* et d'explorer l'objet du point de vue de sa matérialité, de ses contenus, de sa culture ; ce travail n'apparaît cependant que comme un intermédiaire, préalable à une analyse de la façon dont la réalité médiatique du web peut être réinterprétée comme une réalité pour les sciences de la culture. Ce statut d'intermédiaire se reflète également dans ses travaux, parmi lesquels l'un des plus connus est la mise à disposition de l'Issue Crawler, un outil logiciel permettant aux chercheurs de collecter des pages web sans connaissance en programmation, et suivant une méthode qui a été largement documentée pour le monde universitaire. À ce titre, on peut considérer que la deuxième époque, ou deuxième chantier de recherche de la Digital Method Initiative réincarne, avec une contextualisation plus riche et des ponts plus marqués vers les sciences de la culture, le rôle d'équipementier logiciel que peut avoir un Microsoft et que nous avons mis en évidence à travers *The Fourth Paradigm*. De fait, les initiatives de ce type reflètent l'idée qu'il vaut mieux créer des outils, expérimenter, et examiner ensuite les conséquences théoriques et scientifiques de ce que l'on a réalisé, que de vouloir dénoncer ou régler les problèmes théoriques en amont comme on peut le lire dans un certain nombre d'écrits sur les *digital humanities*. À travers ces initiatives, on retrouve non plus un empirisme radical comme dans la mythologie des *big data*, mais un empirisme de l'ordre de la méthode expérimentale de Claude Bernard, qui part des observations et des expériences par essai-erreur, pour aboutir en fin de compte à des connaissances plus générales et plus théoriques.

À l'opposé des conceptualisations plus abstraites à l'œuvre dans le premier âge des *digital methods*, ou des *digital humanities*, cette approche expérimentale est déterminante pour l'apparition de travaux

effectifs fondés sur des traces numériques. Elle peut se faire à partir des sciences de la culture, ou de manière beaucoup plus empirique. En effet, il y a une troisième façon de considérer le web, après l'objet-territoire et l'intermédiaire représentationnel : en considérant comme une brique artéfactuelle, ou une ressource désémantisée. Dans ce cas, il ne s'agit pas en effet de modéliser une représentation du fait humain, mais de s'appuyer sur des traces pour construire un artefact computationnel. Ce statut se présente plus fréquemment en informatique (qu'il s'agisse de chercheurs ou de professionnels), et consiste à considérer le web comme une source opportune de grandes quantités de données servant à la mise au point d'outils. Plus concrètement, dans le cadre de la mise au point de logiciels et de services fondés sur de l'apprentissage automatique *(machine learning)*, de grandes quantités de données d'un genre spécifique sont nécessaires.

Par exemple, dans le domaine du traitement automatique du langage, une tâche fréquemment envisagée est de pouvoir déterminer la tonalité (positive, négative, éventuellement neutre ou mitigée) d'un document. Pour cela l'approche par apprentissage artificiel repose non pas sur la conceptualisation de ce qui est positif ou négatif dans un document, mais sur l'examen massif d'exemples. Ainsi, dans un article intitulé « Twitter as a Corpus for Sentiment Analysis and Opinion Mining », Alexander Pak et Patrick Paroubek (2010) proposent de considérer Twitter comme un corpus pour l'analyse de sentiment. Ils présentent ainsi plusieurs techniques de classification automatique qui sont mises en compétition sur la base d'un même jeu de données (un corpus de tweets). Le type de conclusion qui en ressort ne porte pas sur le sentiment général qui se dégage de Twitter comme objet territoire, ou le fait humain virtuellement accessible à travers Twitter, l'humeur générale de la société, mais sur la capacité de ces différentes techniques à atteindre des scores satisfaisants, et l'intérêt de l'utilisation de Twitter pour ce type de comparaison.

Article	**Citations**[23]	**Statut de la donnée**
« Twitter as a Corpus for Sentiment Analysis and Opinion Mining » (*Ibid.*)	1543	Ressource désémantisée
« What is Twitter, a social network or a news media? » (Kwak et al. 2010)	4619	Objet-territoire
« Twitter mood predicts the stock market » (Bollen, Mao et Zeng 2011)	2221	Représentation

Tableau 5. *Trois exemples de publications incarnant les trois statuts épistémologiques possibles de la donnée issue de Twitter.*

Les trois statuts que nous avons mis en évidence sont conceptuellement distincts, mais ils sont fréquemment amalgamés en pratique. Un même objet, une même source de données, peuvent être envisagés simultanément avec ces trois statuts. Pour illustrer cette ambiguïté, nous avons relevé (**Tableau 5**) trois publications dans des conférences ou revues scientifiques, diffusées à des dates

[23] D'après Google Scholar, consulté le 29 décembre 2016.

proches (2010-2011) et fréquemment citées dans la littérature, qui mobilisent Twitter comme source de données, avec des finalités distinctes.

Dans le premier article, présenté ci-dessus, les données de Twitter jouent le rôle de ressources pour construire un artefact, et auxquelles on n'attribue pas de signification particulière au-delà de la fonction pratique de ressource pour la construction. Elles ne représentent rien, elles servent. Dans le second article, Kwak, Lee, Park et Moon examinent des données de Twitter afin de dégager des généralités sur la façon dont il est utilisé, de manière à mieux comprendre sa nature, et déterminer si les activités qui y ont lieu relèvent davantage de l'information ou de la communication. Ici, Twitter est son propre objet ; le fait que les tweets puissent faire l'objet de traitements quantitatifs leur fait jouer un rôle de preuve des hypothèses sur l'objet. Enfin, dans le dernier article (comme dans celui déjà cité qui visait à prédire les revenus des films à leur sortie en salle), l'objet visé à travers l'analyse de tweets est tout à fait exogène à Twitter, qui lui sert de représentation.

Non seulement une même source de données peut avoir des statuts différents selon l'usage qui en est fait, mais il existe également des cas où le statut de la donnée est ambigu et multiple. Ainsi, le travail mené par Peter Dodds et ses collègues (2015) intègre, comme celui de Pak et Paroubek, une analyse d'un corpus issu de Twitter pour en dégager la tonalité, mais les conclusions vont bien au-delà d'une réflexion sur l'utilisation de ce type de corpus, et porte sur la « positivité » dans le langage. Les tweets ont ainsi tour à tour un statut de brique artéfactuelle et d'intermédiaire représentationnel. Bien souvent, la présentation des travaux qui exploitent des données web ne permet pas de trancher sur le champ visé par lesdits travaux ; il n'y a pas d'inscription claire dans l'un des statuts, ni à l'échelle de l'article, ni d'une section à l'autre de celui-ci. Ce manque de clarté n'est pas à prendre comme une incapacité du lecteur à établir le statut de la donnée dans les publications envisagées, mais bien d'une véritable ambiguïté dans la teneur et l'objet de ces travaux. Dans l'étude du web, bien souvent, on sait ce que l'on analyse, mais pas toujours en vue de quoi : les outils donnent une effectivité expérimentale à ce type d'approche mais pas de cadre interprétatif théoriquement ou historiquement ancré qui permette de donner systématiquement un statut et une signification à la donnée.

Conclusion

En analysant la place donnée au numérique et aux traces dans les sciences de la culture, nous avons établi qu'aucun projet de recherche ne propose de continuité épistémique complète entre les traces numériques, les outils et méthodes et le cadre conceptuel. Nous avons ainsi présenté plusieurs obstacles à cette continuité :

- Certaines sciences de la culture se révèlent incapable de mobiliser de manière opérationnelle le numérique, en raison de freins techniques, méthodologiques, juridiques, économiques, psychologiques ou encore culturels ;

- Face aux freins techniques, elles ne parviennent pas à collaborer avec d'autres acteurs disposant des compétences techniques requises :
- D'autres mobilisent bien des traces numériques, mais qui ne sont pas nativement numériques ;
- D'autres encore mobilisent des traces nativement numériques, mais sont à l'origine de leur production plutôt que d'envisager des traces toujours déjà là de leur point de vue ;
- Enfin, certaines procèdent bien à une conceptualisation des traces numériques comme représentation intermédiaire du fait humain, mais sans compléter cette conceptualisation avec des outils et méthodes explicites.

De ce fait, aucune proposition en provenance des sciences de la culture ne comporte à la fois toutes les composantes requises des sciences computationnelles de la culture et une continuité épistémique entre ces composantes : en particulier, il y a une rupture épistémique systématique entre la trace et ce qui en est fait.

Cette rupture se voit consommée dès lors qu'on envisage une configuration mise de côté jusqu'à maintenant : celle qui consiste à mobiliser des techniques et méthodes exogènes aux sciences de la culture. Comme nous l'avons vu dans les précédents chapitres, le phénomène *big data* n'est pas l'apanage des sciences de la culture, mais de l'informatique universitaire et surtout industrielle. Dans le rapport de pouvoir qu'elles entretiennent avec cette dernière, les sciences de la culture ne sont pas vraiment en position de mettre l'informatique à leur service et de l'inscrire dans leur épistémologie. À quelques rares exceptions près comme le Médialab de Sciences Po, qui se présente comme « composé d'un petit nombre d'universitaires et d'un nombre important d'ingénieurs, qui sont tous considérés comme publiants (qu'il s'agisse d'articles, de logiciels, ou de méthodes) »[24], les sciences de la culture et l'informatique ne forment pas de couple interdisciplinaire à l'intérieur duquel l'instrumentation serait construite en harmonie avec la méthode et le cadre théorique. Il nous faut donc considérer que la rupture épistémique entre les traces et leur instrumentation est un fait inévitable qui ne pourra pas être surmonté, mais franchi ultérieurement, comme on le verra à partir du chapitre VI, par le biais de procédés de reconstitution du sens.

La sortie de l'inscription dans les sciences de la culture ne sera pas sans conséquence : les opérations, les traitements, les outils, n'auront pas de signification *a priori* par rapport à une épistémologie préétablie ; ils ne bénéficieront pas du savoir-faire méthodologique des sciences de la culture. L'alternative à explorer est que ces pratiques computationnelles auront donné naissance à une culture épistémique spécifique, faisant émerger de nouvelles méthodes portant de nouvelles significations. Notre paradigme s'emploierait alors à recombiner les cultures des sciences de la culture et des pratiques computationnelles pour les articuler ensemble et construire une épistémologie commune à

[24] http://www.medialab.sciences-po.fr/fr/about/ (consulté le 21 septembre 2017)

partir d'une compréhension fine de ces deux cultures. Pour explorer cette alternative, nous allons examiner les sciences des données *(data sciences)* considérées comme les pratiques computationnelles nées à partir de la multiplication des traces numériques, et plus spécifiquement le statut de l'instrumentation computationnelle dans ces pratiques.

Chapitre V.
Le statut du calcul dans les sciences des données

La connaissance fondée sur l'observation est généralement médiée par des techniques et des outils qui permettent de rendre ces observations manipulables, intelligibles, interprétables. Dans les sciences de la nature, l'observation à l'œil nu a laissé la place à des instruments qui permettent de voir le lointain, le minuscule, l'invisible, tandis que d'autres instruments permettent de révéler la composition et les propriétés de la matière, d'éprouver la mécanique des objets, ou encore de forcer des signaux à se manifester. Le calcul est l'une de ces techniques épistémiques, qui manipule non pas la nature elle-même, mais sa représentation à travers des données au format numérique.

Dans un contexte de production de connaissances, la conception et la fabrication des instruments s'appuient sur des principes théoriques compatibles avec cette production de connaissances : les instruments de la mécanique, de la thermodynamique, de la chimie, sont fondés sur les mêmes lois de la nature que les branches de la physique auxquelles ils se rapportent. Si leur conception procède également d'un travail de recherche dite technologique, son résultat a un statut d'instrument régi par la problématique scientifique. Comme dans nos hypothétiques sciences computationnelles de la culture, il y a une mise au service de l'outil par la pratique scientifique, qui assure une continuité épistémique entre l'observation, l'instrumentation et le cadre théorique. Cette continuité est une norme historiquement et logiquement constituée d'articulation entre science et technique.

Comme nous allons le voir, les sciences des données mobilisent le calcul comme technique épistémique, mais sans l'articuler avec les sciences de la nature ou de la culture. Elles occupent de ce fait un statut ambigu dans la production de connaissances, suivant lequel on pourrait envisager de les voir davantage comme des arts que des sciences de la donnée. Ainsi, ce sont d'une part des arts fondés sur le calcul, engendrant une réduction du connaissable à ce qui est computationnellement manipulable et calculable, et d'autre part, elles constituent un savoir-faire méthodologique qui donne au calcul un statut spécifique par rapport à ses mobilisations antérieures au service des sciences.

Pour évaluer ce statut spécifique du calcul dans les *data sciences*, et ainsi déterminer ce que la manipulation computationnelle fait à la trace numérique, nous procédons en plusieurs temps : nous

examinons tout d'abord ce que le calcul fait à la donnée et à la connaissance en général, puis plus spécifiquement dans les sciences des données. On verra ainsi que l'intention de mobiliser le calcul suppose un travail de formatage qui permet de rendre les données calculables, mais imprime d'emblée des contraintes et des limites à ce qui est accessible à travers ces données. La mobilisation du calcul à proprement parler engendre également une détermination de ce qui est connaissable en réduisant cet espace à ce qui est calculable, voire programmable. Enfin, cette mobilisation est une activité cognitive qui suscite une certaine tournure d'esprit, un certain rapport à la façon de poser et résoudre des problèmes épistémiques, qui doit être également compté dans les conditions matérielles de possibilité de la production de connaissance à partir du calcul.

Ces conditions de possibilité sont communes à toute mobilisation du calcul. En revanche, l'articulation entre science et technique à l'œuvre dans les sciences des données est singulière par rapport aux pratiques computationnelles dans les sciences de la nature notamment. On présentera ainsi le rapport classique entre science et calcul à travers les notions de modèle, d'expérimentation et de simulation, avant de dégager la spécificité méthodologique (et institutionnelle) des sciences des données dans leur rapport aux techniques computationnelles : les normes de modélisation y sont en effet régies par l'exploration du potentiel du calcul, qui détermine les actions à effectuer sans définition préalable des connaissances visées ou de la compréhension de l'objet que l'on cherche à atteindre. Les sciences des données s'inscrivent ainsi dans un paradigme non plus confirmatoire mais exploratoire, qui ramène les modèles computationnels à des techniques pouvant circuler et être mobilisées indifféremment quel que soit le projet épistémique et la nature des questions qu'il suscite. Ces techniques sont bien souvent les mêmes que dans les sciences de la nature, mais elles y occupent un statut sensiblement différent.

1. Rendre les données calculables

Comme nous l'avons vu au chapitre III, la trace numérique n'a pas de signification préétablie qui permettrait de la soumettre à un traitement calculatoire sans préalablement établir son statut. De surcroît, le geste constitutif qui transforme la trace en donnée par l'attribution d'une charge théorique n'est pas définitif ; il est la première étape de la construction de l'intelligibilité de la trace numérique. Nous montrerons maintenant que ce geste épistémique est doublé d'un geste technique qui rend la trace manipulable au moyen du calcul et en fait une donnée à proprement parler, c'est-à-dire un objet technique inscrit dans un format.

Ainsi, une donnée numérique n'est pas seulement l'inscription d'une observation du réel, investie d'une valeur épistémique actualisable. C'est aussi un objet informationnel qui possède une matérialité technique. La constitution de la trace en donnée est une activité technique aussi bien qu'épistémologique. Cette matérialité constitue la condition de possibilité de tous les traitements computationnels quels qu'ils soient : une donnée non numérique ne peut pas être soumise à de tels traitements, et ceux-ci présentent des limitations théoriques et pratiques. L'espace des connaissances

possibles, productible en analysant cette donnée, est déterminé par cette matérialité informatique, commune à toutes les données numériques en sciences de la nature, dans les sciences des données, ou dans tout autre situation mobilisant ce type d'objet.

De ce point de vue, il ne suffit pas en effet à la donnée d'être chargée d'une valeur épistémique : elle doit s'intégrer dans un format qui la rend manipulable et calculable. De la même manière qu'une liste ou un tableau n'est pas un texte linéaire, la donnée informatique doit être représentée selon certaines contraintes pour sa conservation et sa transmission. Le JSON, le XML, sont des formats classiques de transmission de données informatiques ; l'index et la base de données sont quant à eux des formats emblématiques de stockage. Un jeu de données est décrit et modélisé : de même que les colonnes d'un tableau ont des noms, les types de valeurs stockées sont nommées et leurs relations renseignées.

Quoique ces problématiques puissent sembler fort éloignées des enjeux épistémologiques de l'analyse de données, elles sont les conditions *sine qua non* de cette activité : si le format n'est pas valide, s'il est endommagé, incompréhensible, mal décrit, l'analyste ne pourra réaliser aucun traitement tant qu'il n'aura pas identifié le problème et trouvé une solution pour le réparer, ou obtenu les mêmes données mais sous un format valide. Il s'agit d'une limitation majeure dans la mesure où de nombreuses données qui pourraient se prêter à un traitement computationnel existent sous des formats inexploitables : un PDF est un fichier numérique mais son contenu ne peut pas être manipulé comme une donnée. Dans l'échelle de qualité des données à cinq niveaux proposée par Tim Berners-Lee[25], le PDF est le premier niveau, suivi par le fichier Excel, le CSV, et deux formats propres au web sémantique : le RDF et les *Linked Open Data (LOD)*. Cette typologie de formats est associée à une utilisabilité graduelle : « you get more stars as you make it progressively more powerful, easier for people to use »[26]. D'une certaine manière, une donnée n'est pas une donnée, du moins au sens d'information manipulable et calculable, tant qu'elle ne se matérialise pas dans un certain format qui constitue la condition matérielle de possibilité de son traitement. Bien que l'information représentée soit la même, l'étendue de ce qu'elle peut produire comme connaissance dépend en pratique de ce format :

- la conversion à un format exploitable peut être trop coûteuse pour l'analyste, qui travaille généralement dans un cadre économique et temporel qui n'est pas infini, et estimera excessif l'effort pour la conversion. Par exemple, transformer un PDF de plusieurs dizaines de pages en un tableau CSV est une tâche manuelle longue, fastidieuse et source d'erreurs de retranscription ;
- la conversion à un format exploitable peut tout simplement ne pas être possible du fait d'une incompatibilité entre les représentations des connaissances matérialisées par ses formats. Par exemple, les standards des *Linked Data* intègrent des relations dites sémantiques entre les

[25] http://5stardata.info/en/ (consulté le 16 novembre 2017)
[26] https://www.w3.org/DesignIssues/LinkedData.html (consulté le 16 novembre 2017)

valeurs, qui n'existent pas dans un tableau de données. Un texte n'est pas une liste, qui elle-même n'est pas un graphe ; à chacune de ces représentations correspondent un ou plusieurs formats qui ne peuvent pas être traduits de l'un à l'autre sans perte d'information, voire pas du tout.

En synthèse, le format détermine un espace de possibles au sein du calculable en général : à chaque format correspond un ensemble de manipulations possibles, à la fois théoriques (l'espace de possible tracé par le format) et pratiques (les techniques computationnelles déjà existantes qui y correspondent, et auxquelles l'analyste a accès). En transformant un PDF en feuille Excel, il devient instantanément possible à l'analyste d'appliquer à ses données toutes les formules mathématiques proposées par le logiciel, y compris celles qui n'ont aucun intérêt pour les données en question ; entre le calculable et le calculé, l'espace temporel devient négligeable. Contrairement au mathématicien face à son tableau noir, l'analyste ne manipule plus lui-même mais programme des manipulations dont l'étendue est déterminée par le format ; il ne calcule pas lui-même mais rend calculable. La détermination de l'espace de possibilités computationnelles par le format concerne l'ensemble des données numériques. Que les données soient massives ou non, l'étendue de ce qui est connaissable par le calcul dépend au départ de leur représentation informatique.

Bien souvent, la mise au format est une activité doublement épistémique : d'une part parce que, comme nous l'avons vu, elle détermine l'espace de connaissances possibles à partir des données, et d'autre part parce qu'elle est une procédure où intervient un certain nombre de décisions qui dépendent du contenu informationnel des données et non seulement de leur format. Ainsi, un tableau Excel peut contenir un certain nombre d'informations qui sont lisibles par l'humain mais non par la machine : les commentaires ajoutés aux cellules, leur mise en forme (en gras, en couleur) ne peuvent pas constituer les variables d'une formule mathématique ou logique dans Excel. La mise en forme notamment correspond bien souvent à une sémantique qui 1) n'est pas explicite pour la machine 2) est obscure pour l'analyste pour peu qu'il n'ait pas construit le fichier lui-même, ou qu'un long moment se soit écoulé entre la constitution du fichier et son analyse. De plus, cette mise en forme peut être incohérente, en particulier quand plusieurs personnes ont travaillé sur un même fichier ou que celui-ci est très gros. La coloration d'une cellule en rouge peut signifier qu'elle est importante, ou à supprimer : en l'absence d'autres éléments sémiotiques intégrés au fichier ou en accompagnement de celui-ci, l'analyste qui prépare ses données se verra incapable de leur attribuer une signification stable, et donc de matérialiser cette signification dans un format lisible par la machine. En d'autres termes, ce type d'information ne sera pas calculable, et ne fera donc pas partie du processus de production de connaissances.

Les premiers formats dans l'échelle définie par Tim Berners-Lee sont particulièrement sensibles à ce type d'ambiguïté ou de perte d'information ; néanmoins de nombreux problèmes sont communs à toutes les représentations informatiques de données. Ces problèmes peuvent être de multiple nature :

information manquante, ambiguë, partiellement redondante, contradictoire, illisible, incohérente, etc. En entreprise, il est estimé que le coût du formatage et du nettoyage des données peut représenter jusqu'à 50% du budget d'un projet (Yakout, Berti-Équille et Elmagarmid 2013). Figure emblématique du traitement de données massives, un profil de type *data scientist* passerait, selon les sources et témoignages, entre 60% et 80% de son temps à préparer ses données[27]. Dans certains cas, ce travail de préparation représente même l'essentiel du temps d'analyse des données auxquelles on applique, une fois nettoyées, des traitements élémentaires par comparaison, comme des sommes, des moyennes. Dans ces situations, les difficultés techniques et épistémiques résident non pas dans la sophistication des traitements appliqués aux données, mais dans la constitution de ces données exploitables à partir de données « brutes », pas encore nettoyées ni formatées. La construction de l'exploitabilité des données peut ainsi être vue comme un processus technique et épistémique à part entière, presque suffisant. Des techniques d'apprentissage automatique ont été développées pour automatiser une partie de ce travail (détection et résolution des erreurs) mais le choix de la technique et sa mise au point peut potentiellement être aussi coûteuse en temps que des ajustements manuels à base d'heuristiques. La sophistication de ces techniques d'apprentissage automatique peut elle aussi dépasser celle des traitements computationnels qui seront réalisés par la suite, augmentant leur temps de préparation tandis que le temps d'exécution de ces calculs peut rester négligeable.

Dans le domaine du traitement automatique du langage, la tâche de rendre la donnée calculable se situe à la frontière entre la préparation des données et l'analyse à proprement parler. Un texte n'étant pas, par nature, prédestiné à être soumis à des traitements quantitatifs ou logiques, la constitution des unités sur lesquelles se fera le calcul est en soi un travail d'analyse et d'interprétation, autant qu'une étape de préparation au calcul. Compter les caractères, les mots, les phrases d'un texte, est une opération qui dépend de la façon dont ces unités sont définies : un signe de ponctuation est-il un caractère ? Les formes singulier et pluriel d'un mot doivent-elles elles comptabilisées comme le même mot ? Un retour à la ligne marque-t-il la fin d'une phrase ? Faire du texte une donnée implique de modéliser ces unités dans une double étape de formalisation et de formatage avec codétermination de l'un par l'autre : l'unité « mot » sera à la fois ce qu'on définit et qu'on peut découper techniquement comme telle. Le traitement automatique du langage s'oppose ainsi aux systèmes formels en intelligence artificielle, qui

[27] Voir par exemple :
- https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0
- http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf
- http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#35b787147f75

(consultés le 16 novembre 2017)

> reposent sur l'interprétabilité linguistique de leurs signifiants symboliques pour représenter les connaissances du domaine et sur leur formalité pour être effectif sans qu'il y ait de rapport direct entre les deux. (Bachimont 1996)

À l'inverse, le découpage des mots (ou *tokenization* en anglais) est une tâche simultanément technique et épistémique où le mot est *de facto* défini comme ce qui est découpé comme tel.

La question du format et du formatage ne peut être vue comme une question purement technique, car elle présente des aspects épistémiques cruciaux dans la délimitation du champ du connaissable. Première étape de découverte et de manipulation des données pour l'analyste (à supposer qu'il ne s'est pas déjà chargé de la constitution des données à partir des traces), elle est aussi celle où il se fait une première idée de leur teneur et des traitements qu'il va pouvoir leur appliquer, où il formule, suivant la charge que l'on souhaite donner à cette attitude, des hypothèses ou des préjugés sur les données, qui vont servir de levier et de fil directeur pour le choix des traitements ultérieurs. Sans cette étape de prise en main, les données n'ont pas de sens pour lui, même virtuellement : c'est en les faisant entrer dans un format qu'il les prépare à des traitements compatibles avec ce format et se libère du vertige et de la désorientation qu'évoque la rhétorique des *big data*, en leur donnant sens.

Ainsi, si l'investigation épistémologique est naturellement portée à se concentrer sur les techniques computationnelle de traitement de données, les domaines de l'algorithmique et de l'apprentissage automatique, elle ne peut faire l'économie de l'examen de la façon dont l'information est constituée en donnée, et celui de la façon dont la question en apparence purement technique du formatage et du nettoyage induisent en réalité des choix sur les données effectivement calculables. Tout comme la logique de sélection des traces était, au chapitre III, ce qui les constituait en données pour l'analyste, le formatage est le pendant de cette action pour la machine. La constitution épistémique de la trace se double d'une constitution technique, sans que celles-ci ne soient absolument indépendantes, puisque dans le cadre de l'analyse computationnelle de données culturelles, la faisabilité technique est la condition matérielle de possibilité de la donnée et que les frontières du connaissable sont comprises dans celles du manipulable et du calculable.

2. Le champ du calculable

L'inscription des traces sur un support numérique a en effet des conséquences sur la nature et la portée des manipulations qui peuvent être effectuées pour produire des connaissances : toute connaissance issue de tels traitements est tributaire des limitations techniques et théoriques de ces traitements. D'un point de vue théorique, ce champ du calculable ne paraît guère plus restreint que le champ du connaissable : la distinction pourrait être jugée suffisamment faible pour être négligée. Le domaine de ce qui peut être calculé par un ordinateur dérive de ce qui peut être manipulé dans des termes mathématiques ou logiques. Ce domaine constitue une restriction par rapport aux ambitions

des sciences de la nature par exemple. En effet, les sciences galiléennes se veulent mathématisées, c'est-à-dire qu'elles envisagent ce qui est connaissable comme un sous-ensemble de ce qui est mathématisable. Néanmoins, si certains domaines de la physique contemporaine existent presque exclusivement sous le régime de l'injonction leibnizienne du *calculemus*, la plupart des théories scientifiques reposent sur des connaissances de nature qualitative, et notamment de conceptualisation et de définition des objets. Il ne nous appartient pas de déterminer si cette situation est temporaire, c'est-à-dire d'envisager les connaissances qualitatives comme des connaissances mathématisées en devenir, ou s'il existe des connaissances irréductiblement qualitatives. De ce fait, en l'état des connaissances, le connaissable ne peut être réduit au mathématisable.

Ensuite, il n'y a pas d'équivalence exacte entre le mathématisable et le calculable. Nous savons aujourd'hui que tout ce qui peut être exprimé dans un langage mathématique ou formel ne peut pas forcément être calculé : il existe en effet une classe de problèmes mathématiques dits indécidables, c'est-à-dire qu'on ne peut pas déterminer *a priori* s'il existe une procédure de calcul qui permet d'aboutir à une solution en un nombre fini d'opérations. Historiquement, le problème de déterminer si un problème est décidable a été posé par le mathématicien David Hilbert : c'est le problème *de la* décision, qui porte sur les problèmes *de* décision. La solution à ce problème, établie parallèlement vers 1935-1936 par Alonzo Church et Alan Turing, passe par l'affirmation d'une équivalence entre les problèmes formels envisagés par Hilbert et d'autres formes de problèmes, parmi lesquels on peut prouver qu'il en existe des non-calculables. L'équivalence ainsi soutenue par la thèse de Church-Turing porte d'une part sur le calcul tel qu'il est défini par des systèmes formels comme les fonctions récursives ou le lambda calcul, et d'autre part la notion informelle de ce qui peut être calculé par une procédure effective finie (Pégny 2013). La thèse de Church-Turing affirme ainsi que tout ce qui est calculable au sens informel est calculable au sens des fonctions récursives calculables ; sa converse établit que toute fonction récursive calculable est calculable par une procédure effective, ou autrement dit, un algorithme. De ce fait, toute connaissance mathématisable exprimable par la relation formelle que constitue une fonction récursive calculable peut être considérée comme effectivement calculable en théorie, étant entendu qu'il existe des fonctions mathématiques non calculables, c'est-à-dire des relations formelles qui ne permettent d'aboutir à aucune connaissance par le calcul.

Maël Pégny souligne cependant qu'il y a une distinction assez généralement admise entre la forme algorithmique et la forme empirique de la thèse de Church-Turing : si la première, que nous avons vue, s'intéresse à la notion informelle de calcul comme procédure effective finie quel que soit le substrat matériel où cette procédure est réalisée, la forme empirique s'intéresse aux procédures effectivement calculables par une machine, incluant notamment, mais pas exclusivement, les ordinateurs. En des termes contemporains, elle établit que tout calcul qui peut être effectué par une machine, peut être décrit comme une fonction récursive. Il n'est pas évident que cette thèse soit vraie, mais elle n'a jamais été infirmée en pratique.

L'examen des différentes versions de la thèse de Church-Turing et de ses conséquences se fait idéalement avec beaucoup de nuances et de subtilités dans les affirmations d'équivalence entre les différents types de procédures de calcul : elle cristallise toutes les questions qui préludent à l'émergence de l'informatique théorique puis pratique à partir de problèmes mathématiques et formels formulés au début du XXe siècle. Néanmoins, il nous suffit, pour notre raisonnement, d'établir au moyen de cette thèse qu'il n'y a guère de différence *en théorie* entre les assertions épistémiques exprimables sous forme de fonctions récursives et celles qui sont accessibles par un calcul sur un ordinateur. C'est *en pratique* que le champ du connaissable se restreint, en passant du mathématisé au calculé.

En effet, toute fonction effectivement calculable en théorie ne l'est pas nécessairement en pratique. Les limitations qui interviennent dès lors que l'on considère les aspects empiriques du calcul proviennent de la complexité de la procédure et se manifestent à travers son inscription temporelle[28]. La calculabilité d'une assertion est binaire (il y a des problèmes qui ont une solution accessible à travers une procédure finie, d'autres non), mais sa complexité est graduelle. La théorie de la complexité est un sous-domaine de l'informatique en tant que science théorique du calcul, qui a mis en évidence des catégories de complexité de problèmes, au sein desquelles les problèmes posés sont considérés comme équivalents en termes de complexité : la complexité d'un problème s'établit en théorie mais se matérialise en pratique. Ainsi, les problèmes NP-complets sont une catégorie de problèmes d'une complexité élevée, effectivement calculables en théorie, mais pas systématiquement en pratique. Cette distinction tient au fait qu'un calcul sur un ordinateur est un phénomène physique ancré dans une durée et des contraintes matérielles. Le temps du calcul dépend de la complexité du problème, de l'efficacité de sa solution et de la puissance de l'ordinateur sur lequel il est réalisé. L'informaticien qui conçoit un programme algorithmique dans le but de résoudre un problème concret évolue à l'intérieur de l'espace de négociation défini par ces trois paramètres :

- il peut s'efforcer de ramener un problème appartenant à une classe de complexité élevée à un problème plus simple, moins coûteux en temps et puissance de calcul, notamment au moyen d'une approximation du problème ;
- il peut chercher à améliorer le programme lui-même, en évaluant des alternatives en termes de langage de programmation, de fonctions utilisées, d'ordre des étapes, etc. ;
- il peut enfin, tenter d'accéder à une puissance de calcul plus importante, soit en obtenant un ordinateur plus rapide, soit en distribuant l'effort de calcul sur plusieurs machines.

À l'interface entre ces deux derniers points, il existe par ailleurs une multitude de stratégies d'optimisation de l'adéquation entre le programme (niveau logiciel) et la machine (niveau matériel)

[28] Il existe d'autres limitations à la calculabilité en pratique, comme par exemple les propriétés des valeurs calculées : ainsi, Maël Pégny (2012) souligne qu'on ne pourrait pas effectuer un calcul sur une valeur composée de 10^{250} chiffres car sa taille en nombre de chiffres serait alors supérieure au nombre de particules dans l'univers.

qui visent notamment à répartir le calcul sur plusieurs cœurs du processeurs, à privilégier certaines formes de mémoire par rapport à d'autres (disque dur, SSD, RAM), d'autres types de processeurs (GPU vs CPU), et à optimiser la façon, l'ordre, et les priorités avec lesquelles ces différents éléments physiques sont mobilisés. Ils renvoient par ailleurs à une évolution des pratiques de programmation, entre une première époque qui s'efforce d'améliorer le programme lui-même, de raffiner l'algorithme, son usage de la mémoire, etc., et une période plus contemporaine où l'abondance de la puissance de calcul rend ce travail facultatif.

Bien que ces différentes techniques d'optimisation permettent de négocier la faisabilité pratique d'un calcul, le succès de ces efforts n'est pas garanti : ce qui est calculable en théorie peut se révéler, en fin de compte, non calculable en pratique, ou calculable suivant des approximations ou des reformulations insuffisantes d'un point de vue épistémique. De ce fait, nous avons mis en évidence trois niveaux successifs de limitation que la programmation algorithmique implique par rapport au connaissable en général :

- Tout ce qui est connaissable n'est pas mathématisable (il existe des connaissances de nature qualitative qui ne peuvent pas être formalisées dans un langage mathématique) ;
- Tout ce qui est mathématisable n'est pas calculable en théorie (il existe des problèmes indécidables) ;
- Tout ce qui est calculable en théorie n'est pas calculable en pratique (la complexité algorithmique rend certains calculs impraticables pour une situation technologique et matérielle donnée).

Nous allons maintenant montrer qu'il existe un quatrième niveau, non négligeable, et dont les causes sont davantage psychologiques et métaphysiques que purement épistémologique. Il peut être formulé de la façon suivante :

- Tout ce qui est calculable en pratique n'est pas calculé en pratique.

Ce quatrième niveau, que nous allons analyser avec la notion de « pensée computationnelle », est une restriction supplémentaire par rapport au calculable en pratique, mais qui produit aussi une extension excessive. En effet, on verra au travers de cette notion que deux contraintes supplémentaires doivent être formulées :

- Tout ce qui est calculable ne doit pas forcément être calculé (contrainte morale) ;
- Tout ce qui est calculable ne produit pas nécessairement du connaissable (contrainte épistémique).

À travers cette notion de pensée computationnelle, on verra que la pratique de l'informatique produit deux débordements du champ du connaissable par le calculable, une prédilection pour certaines

formes de production de connaissances, et une croyance dans l'inclusion du connaissable dans le calculable, avec une réduction programmée du connaissable.

3. La pensée computationnelle

Jusqu'à présent nous avons vu en effet les limitations théoriques du calcul, et la façon dont la médiation du calcul implique une réduction de ce qui est connaissable à partir des observables computationnels. Il est crucial d'une part de connaître cette limitation, et d'autre part de garder à l'esprit que toute connaissance ne procède pas du calcul, et qu'il y a d'autres moyens de produire des connaissances. Par ailleurs, ces limitations pratiques s'ajoutent à ces limitations théoriques : tout ce qui est envisageable en matière de traitement computationnel n'est pas mis en œuvre, que ce soit occasionnellement ou généralement. En pratique, la mobilisation du calcul oriente vers certains types d'activités et de projets épistémiques. Ainsi, dans les entreprises en particulier, une très large partie du travail de programmation algorithmique ne consiste pas à trouver une solution à un problème concret mais à optimiser cette solution ; la plupart des problèmes qui se présentent ont des solutions connues, mais pas parfaitement adaptées à la situation. Il y a ainsi une certaine autotélie de la pratique du calcul qui consiste à perfectionner la façon dont elle résout un problème épistémique plutôt qu'à conceptualiser de nouveaux projets. Ce constat pratique favorise une tournure d'esprit computationnelle qui n'est pas sans conséquences épistémiques. Cette tournure d'esprit consiste non pas, comme le souligne Jeannette Wing (2006), en un computationnalisme qui encouragerait à penser comme une machine (et réciproquement, à envisager la pensée comme un calcul), mais en une forme spécifique de créativité :

> Computational thinking is a way humans solve problems; it is not trying to get humans to think like computers. Computers are dull and boring; humans are clever and imaginative. (*Ibid.*)

L'ingéniosité requise pour rendre le calcul possible, pour augmenter son efficacité, pour ronger, parfois milliseconde par milliseconde, sur le temps de calcul d'une tâche qui sera programmée pour s'exécuter des milliers de fois, peut être à la fois bénéfique et problématique. Cette pensée computationnelle induit en effet une prédilection de l'efficience sur la pertinence, et du faisable sur ce qui est souhaitable éthiquement et épistémologiquement parlant. Cette tournure d'esprit est marquée par une recherche d'effectuation du possible qui va favoriser :

- d'une part, la production de connaissances calculables plutôt que d'autres formes de connaissances ;
- d'autre part, l'utilisation de techniques algorithmiques facilement implémentables et économes en temps de calcul indépendamment de la pertinence épistémique de ces techniques et de la signification qu'on peut attribuer à leur résultat.

Le risque extrême en termes de pertinence (mais néanmoins souvent atteint dans les étapes intermédiaires de la mise au point d'une procédure computationnelle) est de parvenir à des résultats calculables mais qui n'ont pas de sens. Moins radicalement, le fait de ramener un problème épistémique à un problème calculable exige une modélisation préalable, c'est-à-dire une simplification du problème, sur laquelle nous reviendrons plus bas. Bien que son propos soit axiologiquement chargé, et illustré avec des caricatures psychologisantes grossières de la figure de l'informaticien, le livre de David Golumbia *The Cultural Logic of Computation* (2009) met occasionnellement le doigt sur des tropismes de l'informaticien qui permettent par exemple d'expliquer le fossé d'incompréhension mutuelle entre chercheurs en sciences de la culture et ingénieurs en informatique évoqué plus haut :

> While philosophers use the term computationalism to refer to others of their kind who believe that the human mind is ultimately characterizable as a kind of computer, here I deploy the term more expansively as a commitment to the view that a great deal, perhaps all, of human and social experience can be explained via computational processes. (*Ibid.*)

De là, on comprend mieux les projets de prédictions de phénomènes socio-économiques à partir de données massives comme ceux que nous avons déjà évoqués : les cours de la bourse ou les succès des films sortis en salle à partir de Twitter, les épidémies de grippe à partir des requêtes sur le moteur de recherche de Google, etc. On comprend mieux également la figure prométhéenne présentée par Pedro Domingos dans son livre *The Master Algorithm*. D'un point de vue épistémologique, la programmation algorithmique trace un domaine de connaissances possibles immense et inédit par rapport à la déjà « déraisonnable efficacité des mathématiques » dans les sciences de la nature (Wigner 1960) ; mais il y a également un aspect psychologique (ou métaphysique), une représentation de la programmation algorithmique par ses praticiens au sein de laquelle on tend vers une superposition de ce qui est connaissable et de ce qui est calculable en pratique. Sans aller jusqu'à l'assertion épistémologique que tout calcul produit des connaissances (car les informaticiens connaissent en revanche la tendance du calcul à produire du non-sens), on peut néanmoins détecter une forme d'injonction morale à calculer tout ce qui peut l'être, mêlé à une forme d'esprit scientifique, au sens d'une inclination à l'expérimentation et au développement d'heuristiques. Il y a ainsi deux extensions illégitimes de la thèse de Church-Turing : celle qui consisterait à envisager tout résultat d'un calcul comme une connaissance, et celle qui ne considérerait comme connaissance que ce qui découle d'un calcul.

La pensée computationnelle des praticiens du calcul ne doit pas être réduite à un tempérament commun fortuit ou qu'on pourrait réduire à des causes sociologiques, dans une perspective externaliste. Elle a un rôle transcendantal dans la pratique de la programmation algorithmique : elle fait en effet partie de ses conditions de possibilités mais détermine aussi ses limites, au sens où la production de connaissances est bornée par les capacités épistémiques de l'agent. Cette pensée computationnelle a également une historicité d'ores et déjà traduite en « roman national » de la programmation informatique, jalonné de figures comme Leibniz et sa combinatoire universelle,

Charles Babbage et sa machine analytique, Ada Lovelace et ses « programmes », et bien sûr les travaux d'Alan Turing.

Dans cet esprit, on reconnaît une influence de la normativité des sciences de la nature, qui confère une légitimité des procédés comme l'expérimentation, la recherche par essai-erreur, bref une exploration plus ou moins systématique du champ des possibles ouverts par le calculable en pratique. L'utilisation épistémique de la programmation algorithmique étant somme toute assez récente (du moins à l'échelle de l'histoire des sciences de la nature), sa pratique est souvent marquée par un esprit de pionnier, voire un esprit de conquête, consécutif à ce qui peut être vu comme la découverte d'un nouveau monde dans le domaine de la connaissance. Dans cette métaphore, l'ancien monde est celui des sciences de la nature, qui servent de référent. Il lègue son état d'esprit, dans une version qui s'apparente fortement à celui de la méthode expérimentale de Claude Bernard, définie de la façon suivante :

> La méthode expérimentale n'est rien autre chose qu'un ensemble de règles sanctionnées par l'expérience et qui ont pour but de prémunir contre les erreurs qui peuvent résulter du maniement des faits et des hypothèses dans l'édification de la science. (Bernard, 1947, cité par Lassègue, 1996)

Dans cette configuration, ce n'est plus l'expérience sur (et dans) la nature, mais le calcul, qui joue le rôle de réalité opposable. Comme on le verra plus bas, l'épreuve du réel n'est plus la pierre qui tombe, le mur dans lequel on se cogne, mais le calcul en pratique dans ce qu'il livre comme connaissance effective de lui-même. L'expérimentation computationnelle peut ainsi assez naturellement être retraduite dans les termes la méthode expérimentale dans les sciences de la nature, comme le propose Jean Lassègue (1996) :

> L'usage heuristique des ordinateurs est semblable à et peut être combiné avec la méthode expérimentale hypothético-déductive traditionnelle en science. Dans cette méthode, on fait une hypothèse sur la base de l'information accessible, on teste expérimentalement les conséquences et on forme une nouvelle hypothèse sur la base de ce que l'on a trouvé ; cet enchaînement est itéré indéfiniment. En utilisant un ordinateur de façon heuristique, on procède de la même manière, sauf que le calcul remplace l'expérience.

Néanmoins, il faut souligner ici que le calcul occupe dans ce cas un statut spécifique d'outil d'expérimentation virtuelle, qui est un exemple particulier d'articulation entre science et technique. De manière plus générale, le calcul possède un double statut de science de l'artéfacture, autonome, théorique et pratique, et de modalité transcendantale empirique d'autres sciences :

- D'une part l'informatique est une **science du possible combinatoire** qui établit de manière théorique et pratique ce qui est accessible par le calcul, celui-ci jouant le rôle de l'objet de la nature jeté devant nous, et dont on cherche à établir les propriétés et les limites pour en fonder l'objectivité. À ce titre, la programmation algorithmique et l'optimisation jouent le rôle d'un

art du possible, c'est-à-dire un ensemble de techniques et de savoir-faire permettant d'établir efficacement des connaissances empiriques relatives à cette science ;

– D'autre part, cet art du possible est un instrument au service d'autres sciences, dont il constitue de ce fait le **transcendantal empirique**, au sens de détermination empirique de l'espace transcendantal des sciences qu'il instrumente. Autrement dit, il délimite empiriquement le domaine de ce qui est connaissable pour ces sciences.

De ce fait, la pensée computationnelle entretient un double rapport aux sciences de la nature : d'un côté, elle peut être vue comme une nouvelle forme d'esprit scientifique imprégnée de la normativité des sciences de la nature, sans pour autant se résumer à cela ; de l'autre, elle devient une composante de l'ethos des sciences de la nature, dans le sens où l'acquisition de cette disposition d'esprit est nécessaire, ou du moins souhaitable, pour l'efficacité de la pratique scientifique à l'heure des ordinateurs. À l'inverse, les sciences de la culture ne bénéficient pas de cette compatibilité naturelle avec la pensée computationnelle, notamment en ce qui concerne ce que nous avons appelé les disciplines à corpus par opposition aux disciplines à terrain (et qu'on pourrait aussi appeler, en utilisant le vocabulaire de Claude Bernard, des sciences expérimentales de la culture). Les normativités des sciences de la nature et des sciences de la culture étant hétérogènes dans l'hypothèse bifurcationniste que nous avons adoptée, les rapports que ces normativités entretiennent chacune avec la pensée computationnelle sont eux aussi distincts.

Dès lors qu'on intègre la pensée computationnelle dans les conséquences de la mobilisation du calcul, et qu'elle est envisagée comme une condition supplémentaire de possibilité des connaissances, le calcul imprime *a priori* au connaissable des contraintes supplémentaires dans les sciences de la culture. Les freins à l'interdisciplinarité entre informatique et sciences de la culture, que nous avons soulevés dans le cadre de notre analyse des *digital humanities*, ne sont dès lors pas seulement des incompatibilités formelles entre des normes scientifiques incommensurables, mais aussi des obstacles d'ordre psychologique, voire métaphysique, entre des conceptions opposées, éprouvant entre elles une aversion réciproque lorsqu'elles sont exprimées dans leurs extrêmes respectifs : une exigence excessive de conceptualisation et de problématisation des objets et des outils d'une part, un injonction effrénée à l'expérimentation et à l'effectuation du possible d'autre part. Dit encore autrement, c'est une incompatibilité pascalienne entre un esprit de finesse porté sur les connaissances qualitatives et compréhensives, et un esprit de géométrie friand d'optimalité et d'efficience. Or, comme l'écrivait Pascal il y a plus de trois siècles, il reste rare « que les géomètres soient fins et que les fins soient géomètres », ce qui fait de la pensée computationnelle une limitation, voire un obstacle, à ce qui est connaissable en pratique et par le calcul, dans les sciences de la culture en particulier. Cette incompatibilité entérine et renforce la rupture épistémique entre la trace numérique et ses traitements computationnels, qui devra être surmontée pour rendre possibles les sciences computationnelles de la culture.

Indépendamment de cette rupture, nous avons, à travers les notions de format, de calculabilité et de pensée computationnelle, analysé la question du calcul et de ce qui est calculable par la programmation algorithmique dans une perspective transcendantale, c'est-à-dire dans le but d'élucider l'articulation qui existe entre le connaissable et le calculable, et plus spécifiquement le domaine du connaissable ouvert par le calcul. Cette perspective rappelle que le calcul est un transcendantal empirique, c'est-à-dire qu'il constitue avant tout les conditions matérielles de production des connaissances qui le mobilisent.

Ce statut est valable quel que soit son contexte de mobilisation. Néanmoins, la mobilisation du calcul dans les sciences des données introduit deux différences principales par rapport à ce domaine. La première, d'ordre technique, exacerbe l'importance de la calculabilité en pratique et de l'optimisation. Des procédés spécifiques, que nous présenterons par la suite, permettent de prendre en charge les caractéristiques techniques des *big data* : leur volume, mais aussi leur caractère incomplet, leur hétérogénéité, le fait qu'elles sont fabriquées et traitées au fil de l'eau, et non en une fois, etc. L'effort requis pour intégrer ces difficultés accentue les prédilections de la pensée computationnelle, et sa recherche d'efficience indépendamment du projet épistémique qui motive le calcul. En entreprise, l'injonction portée par les sociétés de conseil en « *big data* » de se doter des outils pour absorber ces difficultés est même souvent adoptée sans projet épistémique : l'achat de serveurs ou de clusters, la mise en place d'infrastructures distribuées et capables de passer à l'échelle, l'utilisation de bases de données non relationnelles, etc., ne sont fréquemment suivis d'aucune utilisation, car portés par aucun besoin explicite. Ces entreprises, sommées et convaincues de la nécessité d'investir dans « le *big data* », se dotent ainsi de moyens de traitement sans leur associer d'objectif, car elles ne parviennent pas à formuler un projet épistémique à partir de ces données. Dans cette situation, il y a du calculable mais non du connaissable, avec une incapacité à connecter l'un à l'autre. L'exacerbation d'un problème technique (la gestion des propriétés des données) a pour conséquence épistémique de faire échouer à établir, pour commencer, un domaine du connaissable.

La deuxième différence, notamment par rapport à la mobilisation du calcul comme transcendantal empirique dans les sciences de la nature, concerne non pas les propriétés techniques mais le contenu sémantique des données : comme nous l'avons vu en effet, la valeur épistémique des traces numériques relève du domaine des sciences de la culture, tout en procédant d'une rupture avec la continuité épistémique de ce domaine.

Pour mettre en évidence de manière précise le statut spécifique du calcul dans les sciences des données, il nous faut construire un point de repère, montrer par rapport à quoi les sciences des données introduisent un changement. La comparaison avec la mobilisation du calcul dans les sciences de la nature nous permettra de caractériser les sciences de la donnée et de mettre en évidence leur spécificité par rapport aux mobilisations habituelles du calcul dans un projet épistémique, en l'occurrence scientifique. Nous examinons donc maintenant la façon dont le calcul est mobilisé dans

les sciences de la nature dans le contexte de la simulation, du calcul numérique et du traitement de données d'instruments scientifiques.

4. Le calcul dans les sciences de la nature

Il existe aujourd'hui trois modes d'exploitation du calcul dans les sciences de la nature connus de la philosophie des sciences. Bien qu'ils puissent se superposer partiellement, on peut cependant les analyser indépendamment et les désigner de la façon suivante :

- les **simulations numériques** ont généralement pour fonction de proposer une représentation computationnelle, statique ou dynamique, d'un système physique ou biologique ;
- le **calcul numérique** permet de résoudre sur ordinateur des équations mathématiques complexes ;
- le **traitement de données d'instruments numériques** sert à améliorer les qualités (netteté, précision, lisibilité, etc.) des données produites par ces instruments.

En phase avec sa prédilection pour la théorie scientifique plutôt que pour l'expérimentation, la philosophie des sciences classique s'est particulièrement intéressée aux formes et fonctions de la simulation numérique, dont la pratique découle généralement assez directement des problématiques de modélisation scientifique. L'avènement des ordinateurs dans les sciences de la nature permet de renouveler la question du rapport entre théorie, modèle et expérience. Une littérature universitaire de plus en plus conséquente existe ainsi aujourd'hui sur le sujet, héritière des travaux sur le rôle des modèles dans les sciences de la nature.

Cependant, le modèle sur lequel la simulation s'appuie n'est pas la même chose que la théorie. Un modèle scientifique est une représentation d'un phénomène (et non le phénomène représenté, comme le modèle du peintre) qui n'a pas pour fonction de décrire le phénomène de manière exhaustive et dans toute sa complexité, mais d'un certain point de vue correspondant à celui sur lequel on travaille. Ainsi, le graphisme moléculaire bien connus des lycéens, qui représente les molécules par des boules de couleur et des tiges, n'est pas une représentation exacte de la molécule (elles n'ont ni boules pleines, ni couleurs, ni tiges) mais il décrit efficacement le nombre et la nature des atomes, ainsi que les liaisons atomiques dont la molécule est composée.

Il arrive que le point de vue que l'on veut avoir sur l'objet ne permette pas de réduire sa complexité en le modélisant, en l'état des connaissances. Ainsi, en biologie, se présentent fréquemment des objets qui échappent à la modélisation, où dont le modèle le plus simple est l'objet lui-même : le foie, le cerveau, sont des exemples d'objets dont le fonctionnement n'est pas assez connu pour être modélisé et simplifié, de sorte que le seul modèle disponible est l'objet lui-même.

Il est plus facile d'inscrire le modèle dans une épistémologie instrumentale que réaliste : un modèle doit avant tout être fonctionnel du point de vue de la tâche que l'on cherche à accomplir, sans forcément décrire le vrai. Cet aspect fonctionnel lui confère une certaine autonomie vis-à-vis de la théorie dans laquelle il est inscrit. Il peut servir à prouver un aspect de la théorie dont il découle (ou est inspiré) en déroulant ses conséquences, mais peut aussi en être détaché et fonctionner dans le cadre d'autres théories. Il représente ainsi le résultat d'un compromis entre la légalité (le caractère de loi) de la théorie et les contraintes empiriques et pratiques (comme celles de la calculabilité en pratique) qui peuvent amener à des simplifications du modèle initial, ou à l'introduction de modèles équivalents plus faciles à calculer. La négociation peut se jouer notamment entre l'intelligibilité du modèle d'une part, et son exactitude ou son efficacité computationnelle. Par ailleurs, un modèle doit être manipulable, soit littéralement, avec les mains, comme les modèles de molécules en bois ou en plastique que tout lycéen aura connu, soit dans le cadre d'un formalisme spécifique (graphique, mathématique, logique, computationnel) qui définit les opérations légales. De ce fait, un modèle est quelque chose d'opérationnel, qui doit permettre d'expérimenter et explorer les possibilités à l'intérieur des règles formelles définies par le modèle. En ces termes, la simulation numérique est ce qui permet de systématiser et d'automatiser cette exploration sur un ordinateur.

Les pratiques de simulation numérique sont cependant loin d'être homogènes et elles peuvent tour à tour occuper une fonction de validation d'un modèle scientifique, de prédiction, d'expérience ou de support à une expérience scientifique. D'après Franck Varenne (2007), qui distingue dans son vocabulaire la *simulation numérique* (fondée sur le calcul) et la *simulation informatique* (réalisée sur ordinateur), les pratiques de simulations *informatiques* émergent historiquement à partir de deux domaines : les travaux sur les modèles formels d'une part, et ceux sur la simulation *numérique*, où le calcul est réalisé pas à pas par un humain. La simulation repose sur une symbolisation des objets computationnels qui reçoivent une signification avant d'être traités par le calcul. La simulation informatique permet de déduire automatiquement (par le calcul sur ordinateur) les conséquences d'un modèle formel, c'est-à-dire un ensemble d'objets et de relations (notamment mathématiques, statistiques ou logiques) définis formellement. Les pratiques de simulation émergent donc de domaines scientifiques où les mathématiques, le formalisme et le calcul occupent déjà une place importante dans la production et la validation des connaissances. Ce modèle formel doit préalablement être traduit en modèle numérique (ou modèle computationnel) dans lequel les paramètres sont discrétisés et les relations computationnellement calculables (Jebeile 2013). C'est en quelque sorte la version la plus concrète, la plus formalisée et la moins qualitative des différents types de modèles qui coexistent et s'articulent entre eux, à savoir :

- la théorie scientifique, parfois appelée à tort modèle théorique (exemple : le modèle standard) n'est pas un modèle, mais le cadre (et notamment les lois) dans lequel des modèles peuvent être développés ;

- le modèle conceptuel est le modèle scientifique au sens courant, celui qui décrit, dans les termes du domaine scientifique où il s'inscrit, les objets et leurs relations ;
- le modèle formel, qui peut s'exprimer notamment graphiquement, logiquement ou mathématiquement ;
- le modèle computationnel[29], qui décrit les variables, les paramètres, les valeurs, etc.

De manière générale, la fonction du modèle est de servir d'intermédiaire entre la théorie et la situation. De la théorie, il hérite les lois, le régime de vérité et la validité scientifique ; contrairement à elle cependant, il ne décrit pas le réel mais un objet spécifique. Vis-à-vis de la situation, il représente une montée en généralité par conceptualisation et caractérisation de l'objet à travers différentes situations. Cette montée en généralité se fait au prix d'une simplification qui rend le modèle manipulable, et se construit à partir des caractéristiques que l'on veut explorer, de sorte qu'il peut exister plusieurs modèles d'un même objet. Le modèle est une représentation qui sert d'interface entre la théorie et la situation afin de jouer plusieurs rôles : rendre l'objet intelligible à travers une représentation, produire de nouvelles connaissances à son sujet en rendant testables les hypothèses qu'il matérialise, ou encore expérimenter à partir de ce que l'on sait de l'objet.

Si l'élaboration d'un modèle formel du système cible est une condition préalable aux pratiques de simulation, il existe également une pression inverse du modèle computationnel sur le modèle formel (mathématique), voire sur le modèle conceptuel (scientifique). Anouk Barberousse et Cyrille Imbert (2014) ont pu souligner en effet le nombre relativement faible de modèles formels utilisés dans les sciences, et le fait que certains modèles comme les équations de Poisson ou de Lokta-Volterra sont très utilisés. Un même modèle sera utilisé fréquemment au sein d'une discipline donnée mais aussi d'une discipline à l'autre : ce qui est en jeu est en somme une uniformisation des sciences du point de vue des modèles mathématiques qu'elles mobilisent. Anouk Barberousse et Cyrille Imbert suggèrent que ces modèles sont favorisés notamment parce qu'ils peuvent être résolus numériquement (c'est-à-dire *via* un modèle computationnel) et influencent par ailleurs la modélisation conceptuelle des objets scientifiques. En d'autres termes, les auteurs suggèrent une prédilection des chercheurs pour les objets de recherche que l'on peut formaliser mathématiquement et que l'on peut traiter sur ordinateur, au détriment d'autres objets difficilement manipulables formellement et computationnellement : on retrouve ici dans les pratiques de modélisation en sciences de la nature une forme de pensée computationnelle favorisant ce que l'on peut résoudre par le calcul, et restreint le champ de recherche à ce qui est calculable en pratique (dans un temps raisonnable).

La récurrence de certains modèles peut suggérer une interprétation ontologique selon laquelle quelques modèles mathématiques quasi-universels suffiraient à décrire la structure du réel, dans un

[29] Plus précisément, il n'y a pas de modèle computationnel à proprement parler, mais un algorithme ou programme d'un côté, et un modèle implicite de traduction du modèle formel vers cet algorithme de l'autre.

sens platonicien où cette structure serait mathématique. Les auteurs proposent au contraire d'attribuer cette récurrence à des motivations pragmatiques : le fait qu'un modèle puisse être résolu computationnellement, mais aussi l'existence d'implémentations prêtes à l'emploi (notamment sous forme de logiciels) des algorithmes de résolution des modèles, la popularité de certaines approches computationnelles auprès de certains laboratoires ou journaux de recherche, expliquent leur popularité. Cette explication pragmatiste désamorce l'inflation ontologique des modèles mathématiques et computationnels en question, et leur redonne un caractère instrumental où l'outil influe sur les recherches mais ne les gouverne pas nécessairement.

Dans le cas de la morphogenèse des plantes, qu'a beaucoup étudiée Franck Varenne, la simulation à partir du modèle de croissance de la plante permet de valider ledit modèle en déroulant ses conséquences. La simulation joue donc un rôle déductif par rapport au modèle et ses différents niveaux : elle permet de déterminer automatiquement ce qu'il implique. Elle est de nature computationnelle mais son rendu est visuel : son résultat est l'image d'un arbre que l'on peut soumettre à une validation visuelle immédiate, en s'assurant que l'image simulée de l'arbre ressemble à un arbre. Si le résultat visuel de la simulation ressemble intuitivement à un arbre, alors le modèle de croissance implémenté par la simulation est correct. Dans certains cas, la simulation peut être dynamique : le déroulement du calcul consiste en une représentation de la croissance de la plante et permet de visualiser dynamiquement les différentes étapes de la croissance ; le calcul génère une image animée qui correspond au modèle de croissance. Néanmoins, la représentation de la plante n'est pas forcément dynamique : certains modèles produisent une image d'arbre au terme du calcul mais l'inscription temporelle du calcul lui-même n'est pas une représentation (en temps réel, accéléré ou ralenti) de l'inscription temporelle de sa croissance.

Par ailleurs, la simulation elle-même doit être validée : l'ingénieur agronome Philippe de Reffye, étudié par Franck Varenne, a défendu la nécessité d'une validation empirique de l'équivalence entre une simulation informatique et son pendant numérique calculable à la main, ce qui correspond à la fois à un contrôle empirique du calcul et à une construction de la confiance épistémique que l'on a dans la machine. Nous avons vu également que la simulation pouvait être validée intuitivement à partir de ses résultats et de leur cohérence visuelle, sans diagnostiquer les opérations du calcul alors considéré comme une boîte noire. Enfin, la comparaison entre les données simulées et les données expérimentales constitue une forme de validation répandue. Une simulation validée peut avoir de nombreuses applications scientifiques et techniques. Une simulation réalisée à partir de données décrivant le présent permettent de prédire l'état futur d'un système : c'est ainsi que des simulations permettent de calculer des prévisions météorologiques, de déterminer des tendances climatiques ou économiques.

Un statut expérimental est fréquemment conféré aux simulations informatiques : les données qu'elles génèrent sont comparées aux données des expériences menées dans le monde physique. Pour Norton et Suppe (2001, cité par Jebeile, 2016), elles sont une source de données expérimentales et

> les modèles de simulation peuvent fonctionner comme des instruments de mesure fournissant des données sur le monde réel qui font face aux mêmes enjeux épistémologiques que les modèles de données

Dans les cas où il est plus pratique, moins coûteux, etc., de travailler sur les simulations, qu'il n'est tout simplement pas possible d'intervenir sur les objets physiques, ou encore que l'on vise des résultats prédictifs (pour lesquels les données des expériences ne peuvent, par définition, pas encore exister), les simulations se substituent à l'expérience en fournissant des données qui ont, comme les données expérimentales, un pouvoir de preuve ou de rejet d'une hypothèse. Néanmoins, pour que les données simulées aient ce pouvoir de preuve, il faut que le modèle numérique dont elles procèdent soit lui-même validé autrement, sans quoi leur validation respective repose sur un raisonnement circulaire où modèle et données de simulation se justifient réciproquement.

Dans l'ensemble, la simulation numérique entretient des liens forts à la fois avec la modélisation et l'expérimentation. Le calcul y intervient comme une technique qui permet de déduire les conséquences du modèle et de jouer un rôle de substitut ou d'alternative à l'expérience. Les résultats du calcul sont validés d'un point de vue computationnel mais interprétés au niveau du modèle formel et scientifique : la solution computationnelle reste un intermédiaire à la compréhension de représentations plus abstraites.

La simulation apparaît également comme un objet complexe, qui ne permet pas de formuler un point de vue univoque sur son interprétation épistémologique. De plus, elle entretient des rapports multiples avec un autre mode de mobilisation du calcul dans les sciences de la nature, à savoir le calcul numérique, c'est-à-dire la résolution computationnelle d'équations qui représentent généralement des objets physiques. D'une part, d'un point de vue technique, les simulations numériques sont une forme de calcul numérique, au sens où elles prennent la forme de calculs qui sont effectués par des ordinateurs. De même que dans les sciences galiléennes avec une conception instrumentaliste des mathématiques par rapport aux sciences de la nature, les modèles scientifiques reçoivent généralement une forme mathématique, la simulation d'un système physique repose sur la traduction d'un modèle physique en modèle mathématique ou formel puis computationnel. De plus un modèle mathématique peut lui aussi faire l'objet d'une simulation ; comme le suggère Franck Varenne (Varenne et Silberstein 2013) :

> Un calcul numérique de modèle mathématique peut donc bien être conçu comme une simulation de modèle mathématique.

Selon le point de vue qu'on adopte, un calcul numérique peut donc constituer tout ou partie d'une simulation numérique. À l'inverse, il n'est pas rare que le calcul numérique mobilise des techniques de simulation, ou du moins des techniques qui peuvent avoir la double fonction épistémique de résolution d'équation et de génération de données simulées. Ainsi, les méthodes de Monte-Carlo sont souvent présentées comme des méthodes de simulation (Barberousse et Imbert 2013; Jebeile 2013; Boyer-Kassem 2014) mais ce sont également des méthodes stochastiques de résolution d'équation fondées sur la génération aléatoire de données, dont la synthèse fournit une bonne approximation du résultat de l'équation. Une des distinctions claires que l'on peut faire entre calcul et simulation numérique est que les éléments de la simulation, les données générées, ont une fonction de représentation du système cible ; autrement dit, elles ont un sens physique, elles veulent dire quelque chose du point de vue des sciences de la nature.

Ainsi, dans les méthodes de Monte-Carlo appliquées à la physique statistique, les points de données peuvent a) avoir une signification physique (c'est-à-dire représenter des molécules, des états du système, des coordonnées, etc.) ou b) être uniquement des intermédiaires de calcul pour résoudre une équation mathématique (ces méthodes sont, par exemple, souvent utilisées pour la résolution d'intégrales) dont le résultat pourra, par construction, être interprété dans un cadre non seulement mathématique mais aussi physique. Dans le premier cas, les objets mathématiques et computationnels qui constituent la méthode (paramètres, variables, distributions, etc.) doivent être traduits en objets qui ont un sens par rapport au système cible : de cette façon, une même méthode et un même calcul peuvent symboliser un gaz, des particules, un signal électrique, etc. Les techniques de simulation sont généralement mises au point dans un contexte précis : les méthodes de Monte-Carlo sont originellement conçues pour la description de systèmes physiques contenant un nombre élevé de particules, mais sont fondées sur une analogie avec les jeux de hasard (Metropolis et Ulam 1949) et peuvent être utilisées pour décrire une variété virtuellement infinie de systèmes cibles en physique statistique, physique quantique, climatologique, biologie des populations, génomique, etc.

Avec la distinction entre simulation et calcul numérique, on voit que les outils numériques pour les sciences de la nature ont la capacité à circuler de différentes façons entre les disciplines et d'un problème à l'autre : soit au niveau de la simple résolution computationnelle du calcul, soit au niveau du modèle formel dont procède cette résolution. Là encore, c'est la question de l'interprétation et du sens du calcul qui permet de distinguer des fonctions épistémiques et des statuts représentationnels distincts pour une technique de simulation particulière, et également pour caractériser, en les différenciant, les techniques de simulations et les techniques de calcul numérique.

À ces deux mobilisations de la programmation algorithmique dans le cadre des sciences de la nature, il faut ajouter une troisième, celle du traitement de données des instruments de mesure. De nombreux instruments au service des sciences comme les télescopes, les microscopes, les capteurs, les appareils d'imagerie médicale, ont été informatisés, et restituent donc les mesures qu'ils effectuent dans des

formats numériques, qui peuvent d'emblée faire l'objet de traitements computationnels. À noter que bien souvent ces traitements sont, dans les images scientifiques comme dans d'autres domaines, intégrés en réalité à l'instrument. Ainsi, les appareils photos numériques n'enregistrent pas directement le signal de leur capteur mais le compressent, le modifient, l'ajustent, en fonction de paramètres propres à l'appareil et de paramètres définis par l'utilisateur.

Néanmoins, dans le cadre de l'observation scientifique, il y a une suspicion vis-à-vis du traitement de données numériques selon laquelle ces transformations pourraient retirer à la donnée son caractère d'observation du réel et son statut de preuve : le fait de modifier les données des instruments de mesure scientifiques et des capteurs couperait le lien aux phénomènes empiriques qu'elles permettent de construire. Vincent Israel-Jost, qui a travaillé sur le rôle et les modalités de l'observation scientifique, remet en cause l'idée de la supériorité d'une observation qui se voudrait directe et neutre, et montre au contraire que les transformations algorithmiques de la donnée améliorent sa lisibilité et son utilité (Israel-Jost 2016). Il s'inscrit ainsi dans une lignée de recherche plus vaste concernant la place des instruments de mesure dans l'observation scientifique (Hacking 1981; Humphreys 2004; Daston et Galison 2007) et la fonction d'augmentation épistémique *(epistemic enhancers)* qu'ils occupent par rapport à l'observation visuelle directe, qui n'est plus la forme privilégiée d'accès aux phénomènes naturels empiriques.

Dans le domaine de l'imagerie scientifique (qu'il s'agisse d'organismes microscopiques ou de galaxies), le traitement numérique des données générées par les instruments de mesure peut avoir pour fonction d'améliorer la netteté de l'image, de réduire le bruit ou encore de neutraliser les artefacts visuels que peut engendrer la mesure. Il rend ainsi les plus précises, plus exactes, plus lisibles : il ne diminue pas mais améliore au contraire leur caractère observationnel. Dans le domaine de l'imagerie médicale, un traitement computationnel permet également de fabriquer une représentation en trois dimensions à partir d'une série d'images 2D prises par IRM : sans ce traitement, l'analyse de l'image en trois dimensions demanderait un effort de pensée particulièrement complexe. Le traitement numérique de l'imagerie scientifique fournit ainsi dans les sciences de la nature un exemple concret de cas où les données brutes ne sont pas préférables aux données retraitées, voire au contraire, où ce traitement est nécessaire pour rendre les données exploitables. Par ailleurs, il exemplifie également un autre aspect de la préparation de données : le fait que le modèle de l'algorithme de traitement de donnée ne dépend pas de l'objet cible (la galaxie, la mitochondrie) mais de la théorie de l'instrument, de sorte que la même opération de traitement peut être appliquée à des images d'objets complètement distincts. Comme nous l'avons vu dans le cas de la simulation avec l'exemple des méthodes de Monte-Carlo, les traitements computationnels dépendent des théories informatiques et mathématiques sur lesquelles ils reposent (niveau formel), mais sont (sauf exception) indépendants de la théorie scientifique à l'intérieur de laquelle ils visent à produire certaines connaissances.

Si nous avons essentiellement évoqué dans cette section le statut du calcul dans les sciences de la nature, au moins une partie des conclusions auxquelles nous parvenons peuvent être également être appliquées dans les sciences de la culture : il y a une généralité au statut du calcul qui est davantage propre aux sciences *naturalisées* qu'aux sciences *de la nature*. En d'autres termes, c'est au niveau de la méthode d'accès au réel que se définit l'articulation entre calcul et pratique scientifique. Combinée avec la capacité des modèles computationnels à circuler d'un contexte épistémique à l'autre, cette remarque suggère que les pratiques computationnelles issues des sciences de la nature ont également la capacité à circuler au sein de certaines sciences de la culture, en particulier dans les disciplines à terrain.

Qui plus est, la simulation a un attrait supérieur pour ces disciplines : au contraire de la matière ou du vivant, qui peuvent dans beaucoup de cas être étudiés et manipulés *in vitro*, dans l'espace clos du laboratoire, les phénomènes sociaux s'appréhendent généralement « dans leur milieu naturel », là où l'observation ne les modifie pas, et généralement sans intervention. C'est le principe même du terrain, qui repose sur l'observation directe du phénomène en train de se produire. Or, intervenir dans les mondes sociaux soulève des difficultés techniques et des enjeux éthiques que la manipulation de particules ou de champs de force ne pose pas : le recours à la simulation numérique serait une façon pour les disciplines à terrain d'échapper à ces difficultés et à faciliter le travail de recherche, en substituant aux données du terrain les données de simulations numériques.

De fait, la circulation des techniques computationnelles se vérifie déjà au sein de certaines sciences sociales particulièrement mathématisées, comme l'économie, qui utilise par ailleurs des modèles issus de la physique : ainsi le mouvement brownien, qui est un cas de modélisation mathématique du mouvement d'une particule, est également utilisé en finance pour décrire la dynamique d'un cours de bourse. Les systèmes multi-agents sont courants en biologie des populations mais aussi en sociologie quantitative, où des dynamiques de populations humaines sont modélisées et simulées : dans ce contexte les « agents » ne sont plus des organismes vivants mais des individus humains, grâce à une réinterprétation de la fonction épistémique des éléments du modèle de simulation. Le manuel d'introduction à la science sociale computationnelle (Cioffi-Revilla 2014) contient d'ailleurs plusieurs chapitres dédiés à ce type d'approche, qui a également sa revue de recherche (le *Journal of Artificial Societies and Social Simulation).*

Néanmoins, à l'échelle des sciences de la culture, le couple modélisation-simulation fait plutôt figure d'exception, localisées dans quelques disciplines à terrain largement mathématisées, quasi « galiléennes », comme l'économie et la sociologie quantitative. Elle n'a à notre connaissance pas suscité d'initiatives dans les disciplines à corpus qui n'ont pas cette tradition de mathématisation, et ne bénéficient pas par ailleurs de la compatibilité naturelle entre l'esprit des sciences de la nature et la pensée computationnelle. De plus, là où les disciplines à terrain accèdent directement aux phénomènes, les disciplines à corpus mobilisent toujours déjà un intermédiaire de représentation, une

trace qui symbolise un phénomène absent et passé. Il est donc difficile d'imaginer ce que pourrait être une simulation dans une discipline à corpus, en particulier dans celles qui s'intéressent aux connaissances idiographiques. Dans l'ensemble, les sciences de la culture apparaissent comme partiellement compatibles seulement avec les pratiques de simulation numérique, ce qui laisse la place à d'autres mobilisations du calcul.

Il y a ainsi une spécificité de la mobilisation du calcul dans les sciences de la nature à partir des pratiques de modélisation qui peut se résumer en quelques points :

- L'articulation entre science et technique se fait par une instrumentalisation du calcul ; la question scientifique est première. Les problèmes scientifiques conduisent à des solutions computationnelles à travers des modèles de différents niveaux.
- Une culture galiléenne de mathématisation et de modélisation est nécessaire pour voir apparaître des pratiques de simulation au sein d'une discipline scientifique. Dans ces conditions, il n'y a pas ou peu de rupture avec l'épistémologie de la discipline, même si ces innovations impliquent des évolutions techniques et méthodologiques. La modélisation est le point de départ de la mobilisation du calcul.
- Indépendamment de la culture épistémique de la discipline, les modèles formels et les techniques computationnelles ont la capacité à circuler d'une discipline à l'autre, et n'ont donc pas besoin d'être mis au point par un praticien d'une discipline donnée pour être utilisés par celle-ci. Ils peuvent accéder à une certaine autonomie qui leur permet de circuler au-delà du champ du problème scientifique initial.
- Cette capacité à circuler au-delà du problème initial repose sur un processus de décontextualisation puis recontextualisation des objets du modèle computationnel, qui doivent retrouver une signification par rapport au nouveau problème scientifique.

Ces caractéristiques proposent à la fois des clés méthodologiques pour le développement des sciences computationnelles de la culture, et des obstacles à une construction de celle-ci à partir de la culture épistémique des sciences de la culture. Du côté des clés, la capacité de circulation des techniques computationnelles issues des sciences de la nature apporte à la fois des outils concrets pour exploiter les traces numériques, et une logique de décontextualisation/recontextualisation qui permet de leur donner sens. Du côté des obstacles, la nécessité d'une culture de mathématisation et de modélisation suppose un déplacement des sciences computationnelles de la culture en termes de culture épistémique. De ce fait, si la conceptualisation de la trace s'inscrit bien dans cette culture, capable de l'envisager, suivant l'hypothèse bifurcationniste, à la fois sous un angle idiographique et nomologique, la mobilisation des techniques qui permettent de manipuler ces traces pour en dégager une signification se fait dans une autre culture épistémique, confirmant la rupture épistémique entre la trace et sa manipulation. Avant de voir, dans les chapitres qui vont suivre, comment concilier ces deux cultures et surmonter la rupture épistémique entre la trace et le calcul, nous examinons à présent

ce que le calcul fait à la trace dans le contexte des sciences des données, ici envisagées comme une étape ou composante des sciences computationnelles de la culture possédant 1) la culture épistémique requise à travers la pensée computationnelle qui caractérise ses praticiens et 2) construite d'emblée en vue de l'exploitation de traces numériques (sans être, toutefois, conceptualisées comme telles).

En revanche, là où les sciences de la nature mobilisent le calcul comme un instrument, en régissant l'ensemble des objets et étapes de production de connaissance, les sciences des données inversent cette logique : c'est le calcul qui est premier et régit les pratiques. Parmi les niveaux de modélisation qui permettent d'articuler le lien entre la théorie et le réel, seul le niveau du modèle computationnel demeure. Il convient donc de les considérer davantage comme un art du calcul, dont les modèles computationnels invitent éventuellement à une interprétation formelle et conceptuelle, que comme une science à proprement parler où la modélisation est première. Ces arts de la donnée constituent une composante nécessaire des sciences computationnelles de la culture, une condition *sine qua non* à leur réalisation effective, mais leur contribution doit être bornée à l'apport de techniques et d'un savoir-faire relatif au calcul et à la donnée. En ces termes, cette composante peut fonctionner de manière autonome, au sein d'un paradigme méthodologique qui prend la trace déjà constituée comme point de départ et laisse en aval, un espace, un *jeu* pour l'interprétation et le jugement décisif sur la signification de la donnée. Nous examinons maintenant ces sciences des données en tant qu'objet historiquement constitué, pour mettre en lumière la façon dont elles peuvent effectivement fournir la composante computationnelle des sciences computationnelles de la culture, sans outrepasser l'extension de ce domaine.

5. De l'analyse exploratoire à la fouille de données

Au même titre que les *big data* ou les *digital humanities*, les data sciences désignent tout aussi bien un ensemble informel d'objets et de pratiques épistémiques qu'une mythologie associée ; il est donc difficile d'en expliciter le domaine précis. Comme les statistiques, ce sont des pratiques d'analyse de données, mais qui ont plus spécifiquement émergé à la suite de l'avènement des *big data*, donnent au calcul un statut particulier, et s'inscrivent dans un paradigme méthodologique spécifique. Elles reposent, comme on le verra également, sur un creuset de techniques préexistantes issues de lignées techniques historiquement distinctes, dont notamment les statistiques fréquentistes et bayésiennes, la fouille de données *(data mining)* et l'intelligence artificielle. Elles ont ainsi en commun de venir davantage de domaines de recherche technologique, c'est-à-dire de domaines qui interviennent au niveau du modèle computationnel ou formel, que des sciences de la nature. Elles correspondent par ailleurs à des regroupements assez lâches de divers champs de recherche, écoles de pensées, techniques effectivement mises au point, ayant évolué au cours du temps, et appliquées à des problèmes distincts. Nous proposons d'examiner ce que font les data sciences à la trace numérique à

partir de trois cadres historiques, méthodologiques et techniques : celui des statistiques, celui de l'informatique et celui de l'intelligence artificielle.

On verra tout d'abord que les sciences des données s'inscrivent dans un paradigme méthodologique distinct de la statistique classique, en privilégiant un mode exploratoire au mode hypothético-déductif autour duquel cette dernière s'est construite. Ce paradigme est alimenté par des pratiques computationnelles issues d'une circulation institutionnelle et méthodologique entre le monde universitaire et celui de l'entreprise, à travers des individus qui préfèrent travailler à partir d'un problème technologique plutôt qu'à partir d'un cadre institutionnel ou disciplinaire, et se distinguent ainsi de la culture épistémique des statisticiens, mais aussi de l'informatique d'entreprise, par le type de recherche dit « transversale » qu'ils pratiquent. Enfin, le développement des sciences des données conduit à un renouveau de l'intelligence artificielle à partir de ses techniques, qui a pour conséquence de redéfinir son programme et de l'incorporer dans le cadre exploratoire transversal qui les caractérise.

Du point de vue des statistiques tout d'abord, l'apparition des sciences des données se fait plutôt à contre-courant de l'évolution théorique et conceptuelle du domaine. Initialement très descriptive et appliquée à des problèmes empiriques, la statistique est devenue, notamment à partir du début du XXe siècle, un outil de plus en plus formel au service des approches hypothético-déductives dans les sciences de la nature. Néanmoins, cette évolution récente ne révèle en rien la complexité de l'histoire des statistiques, de ses écoles et points de vue. Ainsi, au sein des statistiques modernes, les points de vue fréquentistes et bayésiens correspondent à des interprétations différentes du concept de probabilité : tandis que les fréquentistes analysent les fréquences observées des phénomènes pour en déduire la probabilité, les bayésiens s'intéressent au degré de croyance avec lequel on considère qu'un phénomène peut se produire (Hacking 2002; Barbin et Lamarche 2004). Les statistiques fréquentistes ont connu une activité importante à la fin du XIXe siècle et dans la première moitié du XXe siècle, notamment avec les contributions de Karl Pearson et Ronald Fisher, et se sont structurées en domaine de recherche mathématique à partir des techniques empiriques descriptives développées au XVIIIe et XIXe siècle (Beaud et Prévost 2000; Porter 2003), notamment grâce à leur formalisation par Kolmogorov dans les années 1930 (Lanciani 2011). Les tests d'hypothèse en particulier sont devenus un instrument privilégié de raisonnement inférentiel et de déduction scientifique dans le cadre falsificationniste défini par Popper (Spanos et Mayo 2015). En parallèle, les approches bayésiennes, en disgrâce depuis le XVIIIe siècle, ont connu un regain d'intérêt avec le développement des ordinateurs qui permet d'opérationnaliser ces approches. Les cadres fréquentistes et bayésiens ont tous deux fourni des concepts et des modèles qui ont survécu à leurs différends philosophiques : tandis que les statistiques fréquentistes étaient vues comme déductivistes et les bayésiennes inductivistes, Gelman et Shalizi (2013) ont montré qu'une interprétation hypothético-déductive du bayésianisme est possible, et Mayo et Cox (2009) que le fréquentisme peut quant à lui être inductiviste.

Ces évolutions philosophiques accompagnent les mutations pratiques suscitées par l'informatisation des méthodes statistiques et les nouvelles possibilités induites par le fait que le calcul effectif se fait beaucoup plus rapidement. Les lois de probabilités, l'analyse de la variance, des corrélations, etc., deviennent des outils au service de problématiques et de cadres théoriques variés.

En particulier, l'analyse exploratoire de données développée par Tukey (1977) constitue à la fois une alternative au paradigme hypothético-déductif et une forme de retour à la statistique descriptive, mais avec une nouvelle fonction épistémique. Imaginé avant l'avènement de l'informatique, le paradigme exploratoire de Tukey est sans doute l'une des inspirations majeures de l'analyse computationnelle de données massives : là où la statistique descriptive d'État servait principalement à produire un résumé quantitatif des données, l'approche exploratoire introduit également l'idée de découverte, d'expérimentation, d'identification de régularités mais aussi d'anomalies, bref, de tout ce qui peut mener à l'émergence de connaissances inattendues. La notion d'analyse de données *(data analysis)* est posée dès 1962 en opposition à celle de statistique dans un article intitulé « The future of data analysis » mais c'est en 1977 qu'il publie son manuel d'analyse exploratoire de données. Comme l'écrit John Tukey lui-même :

> This book is about exploratory data analysis, about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights. Its concern is with appearance, not with confirmation.

Les données ne sont plus un échantillon dont on maîtrise la portée épistémique et qui servent à falsifier une hypothèse, mais un donné, au sens de territoire à explorer : peu importe la nature et l'origine des données, qui peuvent être étudiées comme un objet autonome. Néanmoins, l'approche exploratoire se construit à côté, en complément de l'approche « confirmatoire », et non à sa place :

> Once upon a time, statisticians only explored. Then they learned to confirm exactly -- to confirm a few things exactly, each under very specific circumstances. As they emphasized exact confirmation, their techniques inevitably became less flexible. The connection of the most used techniques with past insights was weakened. Anything to which a confirmatory procedure was not explicitly attached was decried as « mere descriptive statistics », no matter how much we had learned from it.
>
> [...]
>
> Today, exploratory and confirmatory can--and should--proceed side by side.

L'ambition de John Tukey est ainsi de réhabiliter une approche qu'il ne présente pas comme son invention, mais comme la forme originaire des statistiques. Il insiste sur le fait qu'il ne s'agit pas d'une approche formelle mais d'un ensemble de techniques globalement élémentaires ; le manuel contient ainsi des techniques de l'ordre de l'astuce pour des tâches très simples comme déterminer le nombre de chiffres d'un nombre sans se tromper, présentées avec un niveau de détail qui va jusqu'à des

instructions comme « posez votre main gauche sur la feuille de papier avec l'index pointant vers la droite ». Ces techniques sont associées à un certain état d'esprit qu'on peut caractériser par la curiosité, l'humilité, la recherche de simplicité. Le manuel de John Tukey a ainsi des accents bachelardien de manifeste d'un « nouvel esprit scientifique », porté par un ensemble de valeurs que doit posséder un analyste de données. L'analyste doit également démontrer une forme d'agnosticité par rapport aux techniques employées, qui n'ont pas de fonction épistémique formellement définie hors contexte. À ce titre, la première technique présentée dans le manuel est remarquable par sa simplicité, son effectivité, et son absence de recours à une machine à calculer. Il s'agit du diagramme tige et feuille *(stem-and-leaf plot)* qui consiste tout simplement à recopier une série de nombres de manière à les regrouper par unité, par dizaine, par millier, etc., en fonction de leur amplitude numérique. Sur la **Figure 4** est ainsi représentée une série de prix allant de 150 à 1895$ (250$, 150$, 795$, 895$, 695$, 1699$, 1499$, 1099$, 1693$, 1166$, 688$, 1333$, 895$, 1775$, 895$, 1895$, 795$), regroupés par centaine. La représentation tige et feuille est construite en écrivant une centaine par ligne (de 1 à 18), puis en répartissant les valeurs dans les lignes correspondantes, en leur retirant les chiffres de 1 à 18 indicateurs de la centaine. Ainsi, 250$ devient 50 dans la ligne 2, 795$ devient 95 dans la ligne 7, etc. Le résultat est similaire à une représentation en histogramme car plus il y a de valeurs dans une ligne, plus cette ligne est longue. Cette visualisation qui peut être réalisée avec un papier et un crayon permet de se rendre compte de la dispersion et de la répartition des valeurs ; elle permet d'identifier les valeurs les plus fréquentes, l'allure générale des données, la présence de valeurs aberrantes. La médiane des valeurs peut aussi être estimée grossièrement mais sans calcul. Elle constitue ainsi une première description des données, indépendante de leur signification, mais à travers laquelle des pistes peuvent être dégagées.

E) UNIT = $1
STEM-and-LEAF
two-digit leaf

1**	50
2	50
3	
4	
5**	
6	95, 88
7	95, 95
8	95, 95, 95
9**	
10	99
11	66
12	
13**	33
14	99
15	
16	99, 93
17**	75
18	95

Figure 4. *Extrait de* Exploratory data analysis*, John Tukey, 1977 (p. 9)*

Un autre aspect important de l'analyse exploratoire de données que met en évidence cet exemple est son exigence de démonstrativité, qui va se matérialiser par l'importance accordée à la visualisation de données :

> The greatest value of a picture is when it forces us to notice what we never expected to see.

Avec l'analyse exploratoire de données, la visualisation acquiert une valeur heuristique. Elle ne sert pas seulement à faire preuve et à communiquer des résultats à ses pairs, mais aussi à rendre les données intelligibles, à surprendre l'analyste et à le faire formuler de nouvelles hypothèses : c'est à la fois un instrument de découverte et de démonstration. L'exploration ne consiste donc pas seulement à parcourir les données, mais à les représenter visuellement de différentes façons de manière à s'en faire une idée et à les connaître de manière de plus en plus précise. En tant qu'outil de découverte, la visualisation de données joue un rôle herméneutique, presque cabalistique, de révélation de ce qui est caché dans les nombres.

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

N = 11
mean of X's = 9.0
mean of Y's = 7.5
equation of regression line: Y = 3+0.5X
standard error of estimate of slope = 0.118
t = 4.24
sum of squares $X - \overline{X} = 110.0$
regression sum of squares = 27.50
residual sum of squares of Y = 13.75
correlation coefficient = .82
$r^2 = .67$

***Figure 5**. Valeurs et représentations statistiques des ensembles de données du quartet d'Anscombe. Tiré de Tufte, Edward,* The Visual Display of Quantitative Information*, p 13.*

Un exemple classique de supplément d'intelligibilité apporté par la représentation graphique par rapport à des représentations statistiques comme la moyenne ou la variance est celui du quartet d'Anscombe (Anscombe, 1973, cité dans Tufte, 2008). Dans cet exemple, quatre ensembles de données (**Figure 5**) composées de deux variables *x* et *y* présentent les mêmes propriétés statistiques (nombre de valeurs, moyenne et variance des *x* et des *y*, corrélation entre les deux, équation de la droite de régression linéaire, etc.). Cependant, lorsqu'on trace le nuage de points de ces quatre ensembles, on se rend compte qu'ils n'ont pas du tout la même allure (**Figure 6**).

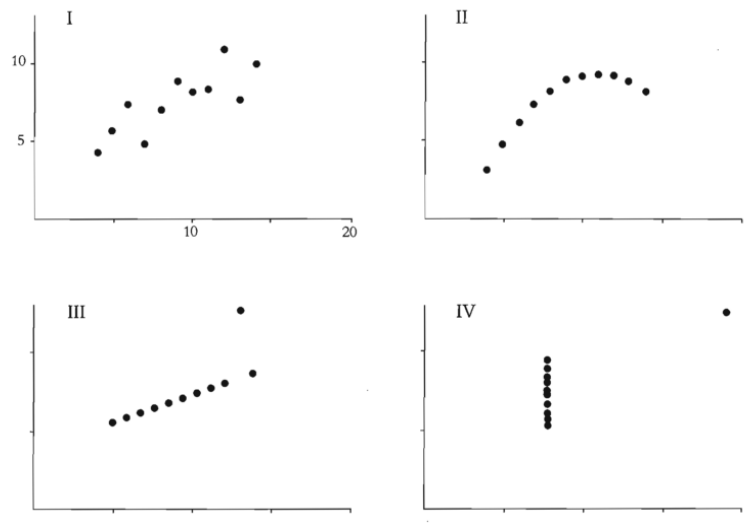

***Figure 6**. Nuages de point des ensembles de données du quartet d'Anscombe. Tiré de Tufte, Edward, The Visual Display of Quantitative Information, p 14.*

Comme nous le développerons au chapitre VIII, la visualisation de données bénéficie de la valeur épistémique attribuée à l'observation directe, et du caractère d'évidence (au sens de ce qui ne fait pas de doute comme au sens de preuve) de ce que l'on peut voir à l'œil nu plutôt que par des représentations purement statistiques comme la moyenne, la variance, etc. Cette évidence de ce qui est représenté visuellement est une vertu épistémique qui transparaît dans notre langage ordinaire, où nous constatons *à l'œil nu* que quelque chose est évident car *cela se voit* et où nous appréhendons le monde qui nous entoure avec notre *regard*. Ce qui peut être *montré* n'a pas besoin d'être démontré :

dans le sens commun, l'observation rend le raisonnement inutile et constitue sa propre preuve. La visualisation de données bénéficie pleinement de la vertu épistémique de l'observabilité. Elle repose ainsi sur le bon sens, le sens commun suffisant à constater ce qui se voit, mais aussi sur l'art de l'observation que développe un œil exercé : l'analyste de données et aussi celui qui sait voir dans les données des indices, des pistes que l'œil ordinaire ne sait pas repérer. La visualisation de données s'inscrit dans une épistémologie de l'œil telle qu'elle a été conceptualisée par Lorraine Daston et Peter Galison dans *Objectivité* (2007), mais aussi à l'art du *jugement exercé* développé dans le même ouvrage comme l'une des vertus épistémiques fondatrices de vérités scientifiques.

Néanmoins, le savoir-faire à l'œuvre dans l'analyse exploratoire de données ne repose pas seulement sur la vision. C'est un jugement exercé mais aussi une manipulation exercée, un jeu entre voir et faire. La combinaison de l'intuition, de l'expérience, et de la maîtrise des outils statistiques que l'analyste a à sa disposition, sont nécessaires pour instrumenter l'inventivité avec laquelle il regroupe, séquence et représente les données pour identifier des angles intéressants. En l'absence d'une méthode formelle (ou informelle, mais qui puisse *a minima* être enseignée), l'analyste doit inventer des solutions méthodologiques et graphiques à chaque nouvel ensemble de données en puisant dans la boîte à outil proposée par John Tukey. Cette inventivité par laquelle il crée de l'intelligibilité se double d'une sagacité, par laquelle il doit déceler le non-sens que peut introduire une visualisation. Une représentation peut en effet être trompeuse : pour peu, par exemple, que l'outil de visualisation utilisé tronque automatiquement l'axe d'un histogramme, l'écart entre deux séries de données paraîtra disproportionné, créant une impression factice de sens. En l'absence de normes comme la valeur p dans les tests d'hypothèses, cette sagacité consiste également à pouvoir distinguer ce qui est significatif de ce qui ne l'est pas, ce qui relève de l'information ou du bruit statistique.

L'analyse exploratoire de données se présente donc comme un art, à la fois comme savoir-faire non formalisé, préscientifique, et dont l'exécution dépend de la qualité de l'analyste, et comme un travail esthétique à travers lequel la visualisation de données joue un rôle rhétorique de séduction et d'argumentation. En tant que savoir-faire l'art de repérer les pistes et les indices est un aspect constitutif de l'analyse exploratoire de données que l'on peut relier au paradigme indiciaire à l'œuvre dans les sciences de la culture et que nous avons déjà évoqué. Comme Ginzburg, John Tukey mobilise la métaphore du détective et de l'enquête policière :

> A detective investigating a crime needs both tools and understanding. If he has no fingerprint powder, he will fail to find fingerprints on most surfaces. If he does not understand where the criminal is likely to have put his fingers, he will not look in the right places. Equally, the analyst of data needs both tools and understanding. It is the purpose of this book to provide some of each.

Le manuel de John Tukey se présente ainsi comme un manuel d'investigation pour analyste de données. Celui-ci se fonde, comme l'écrit Ginzburg, « sur des détails qui échappent à la plupart des gens ». Les techniques d'analyse exploratoire de données constituent l'outillage de ce travail

d'observation : comme le luminol permettant de révéler ce qu'on ne voit pas à l'œil nu, elles mettent en évidence ce qu'on ne voit pas en regardant les données elles-mêmes, mais une représentation graphique donnée. La médiation de la visualisation de données se combine au jugement exercé de l'analyste de données pour repérer des indices. Ce « détective » singularise des éléments qui présentent des suppléments de sens par rapport à l'ensemble : irrégularités, anomalies, éléments surprenants par rapport à des connaissances préétablies. Les techniques statistiques et graphiques pour représenter synthétiquement les données ne servent pas tant à les décrire dans l'ensemble, qu'à fournir un cadre de comparaison pour mettre en évidence des variations. Les moyennes, les régularités, sont les repères, le cadre de référence à partir duquel l'analyste recherche de la différence statistique.

De ce fait, rien de ce qui est mis en évidence avec l'analyse de données ne ressort automatiquement : les techniques de l'approche exploratoire instrumentent un analyste dont le jugement exercé est la clé de compréhension des données. Néanmoins, à l'inverse du détective ou du critique d'art de Ginzburg, l'indice du détective à la Tukey est déjà une quantité singularisée comme signifiante, et non une unité, un objet singulier : ce n'est pas une seule branche cassée par la bête que traque le chasseur, mais un nombre exceptionnel, statistiquement signifiant, de branches cassées qui peuvent être l'indice, par exemple, de l'agitation d'une population d'animaux. L'indice de la statistique exploratoire est une quantité qui dérive d'une norme au sein de données, plus qu'un élément qui se détache de l'arrière-plan du réel. De plus, la forme de restitution par excellence du paradigme indiciaire est le récit, l'argumentation, c'est-à-dire dans les termes de Ricœur (1983) une synthèse temporelle de l'hétérogène ; à l'inverse, celle de l'analyse exploratoire est initialement une synthèse spatiale ou topologique, reposant sur la rhétorique graphique de la visualisation de données, bien qu'une mise en récit des données puisse être proposée par la suite, comme on le verra au chapitre VIII.

Le paradigme exploratoire développé par John Tukey est antérieur à l'informatisation de l'analyse de données et de la statistique. En parallèle des transformations suscitées par le développement de l'informatique dans les sciences de la nature, il s'est néanmoins lui aussi construit une forme computationnelle par la suite : s'il faut ainsi retracer la généalogie de l'esprit des sciences des données, c'est chez Tukey, par l'intermédiaire de la fouille de données, qu'il faut aller chercher une origine. L'approche exploratoire a ainsi trouvé des applications dans l'univers de l'informatique d'entreprise sous la forme de la fouille de données *(data mining)* et de la découverte de connaissances *(knowledge discovery)*. Bien que la fouille (ou exploration) de données ne soit pas forcément présentée comme une technique exclusivement industrielle, ses principales applications sont le marketing, la finance, la détection des fraudes ; les exemples cités dans la littérature portent généralement sur l'analyse de données relatives à des consommateurs (Fayyad, Piatetsky-Shapiro et Smyth 1996). L'exploration des données est conditionnée par l'existence de bases de données structurées en entrepôts de données

(data warehouses) qui suivent des modélisations conventionnelles et servent à agréger les données « métiers » que produit et/ou dont dispose l'entreprise : elle émerge du développement de l'informatique décisionnelle *(business intelligence)* dans les années 1990 et a pour fonction de faciliter la prise de décision dans les entreprises.

Le terme d'extraction de connaissance, ou plus exactement de *knowledge discovery in databases* est attribué à Gregory Piatetsky-Shapiro et défini comme « l'extraction spécifique d'information implicite, précédemment inconnue et potentiellement intéressante, à partir de données »[30] ; il s'agit notamment d'identifier des régularités *(patterns)*, des sous-ensembles d'éléments similaires, des anomalies. Ce domaine s'appuie sur les notions d'exploration, de découverte, et sur l'idée que l'objet étudié est la donnée elle-même et non ce qu'elle représente. C'est une approche guidée par les données *(data-driven)* dont la connaissance doit être le produit final, et à l'intérieur de laquelle la fouille de données n'est qu'une étape parmi d'autres telles que la sélection, le prétraitement, et l'interprétation *(Ibid.)*. Comme l'analyse exploratoire de données de John Tukey, la découverte de connaissances est une approche générale davantage adossée à un but et un état d'esprit qu'à des techniques spécifiques. Elle constitue le paradigme méthodologique des sciences des données.

À l'intérieur de ce paradigme, l'analyse de données à proprement parler s'appuie indifféremment sur des techniques issues de l'informatique d'entreprise, de la recherche en sciences de la nature et de la culture, en statistiques ou en intelligence artificielle, ces catégories étant elles-mêmes des simplifications des champs de recherche fondamentale ou appliquée aboutissant à ces techniques. Nous examinons maintenant, à travers les principales techniques des sciences des données, le statut de recherche transversale qui rend possible cette indifférence aux cadres théoriques historiques et cette facilité de circulation entre projets épistémiques.

6. Le creuset institutionnel et théorique des sciences des données

Les différentes techniques computationnelles mobilisées en exploration de données ont fait l'objet de nombreuses tentatives de classification en fonction de leur origine, de leur fonction ou de leur fondement théorique. Ainsi, le manuel *Mining Massive Datasets* (Rajaraman et al. 2011) organise les algorithmes de fouille de données en fonction des opérations faites sur les données (regroupement, recherche d'anomalie, recherche de similarités, etc.) aussi bien que de nature de données (données en flux, grands ensembles) ou encore de problématique épistémique (recommandation, publicité, analyse du web), celles-ci pouvant également être de niveau différent ; ainsi la recherche de similarité fait partie des techniques permettant de générer des recommandations qui peuvent être appliquées à des problématiques de ciblage publicitaires. De même, le manuel *R and Data Mining: Examples and Case*

[30] 'Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.' (Frawley, Piatetsky-Shapiro, & Matheus, 1992)

Studies (Zhao 2012) propose des regroupements à la fois en fonction du type d'opération effectuée et de la lignée technique de l'algorithme : ainsi, le *clustering*, les arbres de décision et la régression sont trois genres de techniques de classification, mais font l'objet de chapitres différents. À un niveau plus conceptuel, Nicholas Diakopoulos (2015) propose d'organiser les techniques des sciences des données selon le type de décision prise par l'algorithme : priorisation, classification, association, filtrage.

En réalité, du point de vue des opérations faites sur la donnée, nous identifions trois grandes familles :

- la **hiérarchisation**, qui attribue un rang ;
- la **classification**, qui attribue une catégorie ;
- le **regroupement**, qui attribue des éléments similaires.

Le filtrage au sens de Diakopoulos est une forme de classification binaire qui découpe un ensemble entre filtré et non filtré. L'association est une forme de regroupement, tandis que la hiérarchisation inclut ce que Diakopoulos appelle priorisation. Néanmoins, lorsqu'on examine la question du point de vue du problème épistémique posé, ces différentes familles peuvent être mobilisées alternativement ou de manière complémentaire. Par exemple, un moteur de recherche hiérarchise des résultats de recherche, mais il peut également s'appuyer sur l'identification de termes similaires aux termes de la requête, ou renvoyer tous les résultats correspondant à la même catégorie thématique que les termes de la recherche. Si ces familles constituent des clés d'entrée dans l'univers des algorithmes d'analyse de données, aucune tentative de classification de ces algorithmes ne sera satisfaisante du fait de leur circulation constante et de leur polyvalence en termes de problème technique et épistémique, ou en d'autres termes, de leur capacité à appartenir à toutes et aucunes de ces familles en même temps. L'origine historique, la communauté de recherche, le cadre théorique, ne déterminent en aucun cas les usages et fonctions de ces techniques, qui y sont d'une part indifférentes, et circulent d'autre part d'un cadre à l'autre.

Au plus bas niveau, on dispose de bibliothèques *(libraries)* ou d'implémentations spécifiques qui permettent d'utiliser effectivement un algorithme existant, comme par exemple l'algorithme APriori développé par Rakesh Agrawal et Ramakrishnan Srikant (1994). L'algorithme APriori est défini comme un algorithme de règles d'association, c'est-à-dire une technique qui permet d'identifier des éléments fréquemment associés. Historiquement, il est connu pour son application dans la grande distribution où il permet de repérer les produits fréquemment achetés ensemble en analysant les historiques des caisses de supermarché. Le fait d'être fréquemment associé peut être considéré comme une manière de mesurer une certaine forme de similarité entre éléments, ou du moins une manière de regrouper des éléments suivant une certaine logique systématique. Il existe par ailleurs d'innombrables techniques de regroupement automatique d'éléments dans la famille des algorithmes de classification non supervisée. Les notions de classification supervisée et non-supervisée viennent pour leur part du domaine de l'apprentissage automatique *(machine learning)*, lui-même considéré

comme une application ou un prolongement du projet général de l'intelligence artificielle. Bien que les règles d'association et la classification non supervisée soient deux familles d'algorithmes techniquement distinctes, fonctionnant selon des principes différents, elles peuvent être rapprochées du point de vue de leur fonction, c'est-à-dire réaliser mettre en évidence des similarités entre éléments, et ce quels que soit le sens initialement associé à ces éléments (produits, personnes, mots, etc.).

À travers l'exemple du regroupement automatique, on voit ainsi qu'il y a, du point de vue des fonctions épistémiques mais non des origines historiques, de fortes superpositions entre des techniques de fouille de données mises au point dans l'industrie et des techniques d'apprentissage automatique développées par des chercheurs. De plus, comme on l'a évoqué dans notre analyse de la mythologie des *big data*, le traitement de données massives remet en cause la pertinence de cette dichotomie industrie/recherche. On voit en effet émerger des acteurs qui échappent à cette dichotomie et fondent dans cette liberté la richesse de leur travaux. Le sociologue des sciences Terry Shinn (2000) a ainsi conceptualisé la notion de « régime transversal » pour décrire des praticiens dont le travail s'organise autour d'un projet plutôt qu'autour de disciplines ou d'organismes spécifiques. Ces praticiens peuvent être tout à la fois universitaires, ingénieurs, entrepreneurs, consultants, en échappant en même temps à toutes ces catégories. Avec le régime disciplinaire, qui maintient l'appartenance à une discipline, et le régime transitaire, dans lequel le praticien « traverse les frontières de sa discipline » pour emprunter à des disciplines voisines, le régime transversal est l'un des trois régimes de recherche mis en évidence par Terry Shinn. Le régime transversal peut notamment prendre la forme d'une recherche « technico-instrumentale ». Ce type de recherche quant à lui :

> tourne autour de pratiques concrètes qui incluent la conception, la construction, le bricolage sans fin et l'analyse destinés à justifier les principes de base des instruments, l'adaptation pour améliorer leurs performances, les recherches et contrôles destinés à définir dans quelle mesure un instrument générique peut être généralisé, les tests et modifications permettant de vérifier si les principes de généralisation tiennent et le transfert de l'instrument vers des usagers pour l'adoption finale et la mise en œuvre.

Marquée par l'opportunisme et le rejet d'allégeance à un organisme en particulier, la figure du chercheur technico-instrumental passe fréquemment de l'industrie à la recherche et vice-versa ; sa production intellectuelle prend la forme d'instruments concrets, de publications scientifiques, de rapports, de brevets, de conférence. L' « artisan » de la recherche technico-instrumentale s'intéresse à la conception d'instruments plus qu'à une activité théorique, et s'efforce de produire des outils génériques, qui peuvent avoir de multiples applications. Aujourd'hui, cette figure pourrait vraisemblablement s'apparenter à certains types d'entrepreneurs influents qui interviennent indifféremment devant un public universitaire, professionnel, ou amateur, avec une capacité à se médiatiser sans être placés dans une catégorie.

Les concepts développés par Terry Shinn sont particulièrement utiles pour décrire les acteurs qui conçoivent et développent les algorithmes du traitement de données massives. Pour comprendre l'origine de ces techniques, il nous faut non seulement analyser les circulations entre les institutions définies par la sociologie des sciences classique, mais également s'abstraire de ces catégories pour comprendre les conditions d'émergence de ces techniques. De la même manière que les techniques elles-mêmes circulent d'une institution à l'autre, les acteurs aussi naviguent entre des contextes institutionnels changeants. À ces conditions matérielles correspond un cadrage (ou absence de cadrage) théorique qui a une influence directe sur la conception et les usages prévisionnels de ces techniques.

Si l'on examine maintenant les algorithmes les plus « influents » en fouille de données, on se rend compte en effet qu'ils sont presque tous le résultat de trajectoires transitaires ou transversales. Pour illustrer ce constat, nous nous référons à une enquête (Xindong Wu et al. 2008) présentée lors de l'une des principales conférences de recherche en fouille de données, la International Conference on Data Mining (ICDM) de 2008 organisée par l'IEEE. Cette enquête s'appuie sur les contributions des gagnants de plusieurs trophées, l'ACM KDD Innovation Award et l'IEEE ICDM Research Contributions Award. L'IEEE et l'ACM sont les principales organisations de chercheurs en informatique, ici mobilisées dans le domaine de la découverte de connaissance et de la fouille de données. Ces gagnants ont identifié un certain nombre de techniques. Celles qui avaient moins de 50 citations dans des articles de recherche ont été éliminées. La liste a ensuite été soumise au vote d'une multitude d'acteurs du domaine (organisateurs et participants à des conférences de recherche) dont les votes ont concordé. Si cette enquête ne permet pas forcément de dégager les principales techniques du traitement de données massives en général (et ne prend pas en compte des acteurs uniquement industriels), elle donne néanmoins une bonne idée de techniques reconnues dans le domaine.

Dans le « top 10 » mis en évidence par cette enquête, on retrouve l'algorithme APriori que nous avons déjà évoqué, ainsi que d'autres familles de techniques. La moitié de ce top 10 est constituée de techniques de classification supervisée (C4.5, CART, SVM, kNN, Naive Bayes) qui constitue en effet le problème le plus courant en apprentissage automatique. Sont également présentes des méthodes de classification non supervisée (k-Means), d'estimation de paramètres (EM), de tri (PageRank) et de boosting (optimisation de classifieurs : AdaBoost).

Comme l'algorithme APriori incarnait un exemple de circulation d'un problème industriel vers des usages plus génériques, le k-means est un exemple de trajectoire d'une technique issue de la recherche en sciences de la culture. Il est en effet l'œuvre d'un psychologue, James McQueen, intéressé par la modélisation mathématique des activités humaines ; dans l'article décrivant sa méthode, il présente ainsi plusieurs applications sur des variables à la fois comportementales et textuelles issues d'observations en psychologie sur des populations d'étudiants (MacQueen 1967). Enfin, la circulation disciplinaire d'une technique peut être faite par ses concepteurs eux-mêmes : une technique

d'indexation documentaire aujourd'hui banale, l'analyse sémantique latente *(LSA)*, qui a fait l'objet d'un brevet en 1988 et d'une publication deux ans plus tard dans une revue de sciences et technologies de l'information (Deerwester et al. 1990), a été ultérieurement remaniée par ses concepteurs pour des usages en psycholinguistique (Landauer et Dumais 1997), avec l'ambition théorique de mieux comprendre l'acquisition du langage chez les enfants. L'article correspondant est d'ailleurs publié dans une revue de psychologie, la *Psychological Review*.

C'est aux laboratoires Bell, où ont également travaillé Turing, Shannon et Tukey, que Vladimir Vapnik et ses collègues ont mis au point les machines à vecteurs de support (SVM), qui font partie du top 10 de l'ICDM. Les SVM sont extrêmement populaires en fouille de données du fait de leur simplicité, de leur efficacité, et de la qualité des résultats qu'elles apportent. Elles ont ainsi rapidement supplanté les techniques à base de réseaux de neurones au moment de leur apparition. La migration institutionnelle de Vapnik (qui travaille aujourd'hui chez Facebook) de la recherche à l'industrie se reflète dans ses travaux théoriques : en 1998, il propose en effet la théorie statistique de l'apprentissage (Vapnik 1998) qui apporte un fondement statistique à l'apprentissage artificiel et assure une continuité théorique entre ces deux cultures souvent opposées.

Un autre exemple d'individu qui incarne une synthèse entre ces deux cultures est Leo Breiman, chercheur en statistique devenu consultant puis redevenu universitaire. Il ainsi un parfait exemple du chercheur et artisan technico-expérimental décrit par Terry Shinn. Breiman est plus connu dans la communauté informatique pour avoir mis au point les techniques de *random forests*, étonnamment absentes du classement de l'ICDM. Dans un article de recherche sur les « deux cultures » en modélisation statistique (Breiman, Cox et Breiman 2001), il décrit l'état d'esprit des praticiens de l'apprentissage automatique avec lesquels il a pu travailler dans l'industrie lorsqu'il était consultant, et la communauté naissante d'ingénieurs et chercheurs en informatique et physique qui travaillent sur ce type d'approche à partir des années 1980. De ce fait, Leo Breiman est non seulement un acteur emblématique de la recherche techno-instrumentale sous régime transversal à l'œuvre dans les sciences des données, et un contributeur marquant de ce type de recherche, mais aussi un observateur, quasiment un analyste, des conditions sociales et épistémiques de l'avènement des sciences des données à partir de multiples lignées techniques et disciplinaires.

Dans son article, Leo Breiman regrette le conservatisme de la culture « statistique » tel qu'il l'oppose à cette culture naissante de « modélisation algorithmique ». L'opposition entre ces deux cultures est manifeste à travers la réception de l'article, commenté avec virulence par des représentants éminents de la culture statistique qu'il dénonce, comme David Cox et Brad Efron. Que cette opposition soit philosophique, culturelle, générationnelle, institutionnelle ou un mélange de ces aspects, n'est pas évident à la lecture des commentaires, mais elle y apparaît avec une force qui dépasse celle d'un échange d'arguments sur des techniques de calcul.

Ces deux cultures coïncident ainsi partiellement avec l'opposition entre pratiques universitaires et pratiques industrielles, et avec certaines prédilections de part et d'autre. Elles renvoient également à des techniques qui leurs sont plus spécifiques. Si elles suscitent d'intenses débats, c'est sans doute aussi car les deux cultures présentent de part et d'autres de nombreux points communs, éventuellement décrits avec des termes différents. Robert Tibshirani, professeur de « data sciences biomédicales et statistiques » à l'université de Stanford, a ainsi mis en regard un certain nombre d'éléments de comparaison entre apprentissage automatique et statistiques dans un but de clarification pour ses étudiants **(Figure 7)**, suggérant qu'un même algorithme peut être interprété dans le cadre de l'apprentissage automatique comme dans celui des statistiques. En effet, si ces deux cultures peuvent être opposées de manière si précise, c'est parce qu'elles relèvent d'une tâche commune, l'analyse (désormais systématiquement) computationnelle de données. C'est au niveau du cadre conceptuel, de la culture épistémique et de la communauté de recherche que se joue la différenciation entre ces deux cultures.

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

***Figure** 7. Notes de cours de Robert Tibshirani, professeur de « data sciences biomédicales et statistiques » à l'université de Stanford.*

L'article de Leo Breiman cristallise ainsi un rapport de forces entre deux cultures épistémiques au tout début des années 2000, avant l'émergence du phénomène *big data*. Aujourd'hui, l'articulation entre ces deux cultures s'est complexifiée :

– certains statisticiens considèrent tout simplement que les data sciences sont un terme à la mode pour désigner les statistiques, et par conséquents qu'ils sont data scientists[31] ;

[31] http://simplystatistics.org/2011/11/22/data-scientist-vs-statistician/(consulté le 16 novembre 2017)

- certaines techniques statistiques comme la régression logistique sont fréquemment décrites comme des techniques d'apprentissage automatique[32] ;
- d'autres statisticiens se sentent menacés de disparition par la part croissante de leur travail qui peut être automatisée[33] ;
- les praticiens des data sciences n'opposent généralement pas sciences des données et statistiques, mais considèrent les statistiques comme une compétence indispensable à un bon data scientist[34].

Par ailleurs, si les techniques et le savoir-faire statistiques sont effectivement absorbés par les sciences des données, la culture épistémique propre aux statistiques ne semble pas plus soluble dans celle des sciences des données que dans celle de l'apprentissage automatique, avec laquelle elle ne coïncide pas tout à fait. L'opposition tracée par Breiman persiste du point de vue de la « philosophie de la donnée » pratiquée par ces deux cultures.

Une manière de comprendre l'incommensurabilité de ces deux cultures est de les comparer à l'incommensurabilité des paradigmes scientifiques chez Kuhn. Statistiques et sciences des données correspondraient alors à des cadres théoriques propres et incommensurables, à l'intérieur desquels sont développés des modèles formels et computationnels qui peuvent quant à eux circuler d'un cadre à l'autre. Ce ne sont pas les modèles mais les théories qui sont incompatibles. Ainsi, en tant que cadre théorique, la physique newtonienne et la relativité générale sont incommensurables : elles correspondent à des représentations du monde qui se contredisent mutuellement. Néanmoins, au niveau de ses modèles, la physique newtonienne n'est pas invalidée, mais perfectionnée par la relativité générale. Ses équations sont moins exactes, mais constituent des approximations souvent suffisantes pour représenter un système cible ; par exemple, prendre en compte la courbure de l'espace-temps pour modéliser l'écoulement de l'eau dans une baignoire est une sophistication qui complexifie le modèle analytique sans apporter de précision intéressante. Dans les sciences de la nature, deux théories peuvent être incommensurables tout en laissant la possibilité d'une traduction des modèles de l'une dans l'autre.

De la même manière, on peut considérer que les statistiques et les sciences des données constituent deux théories mathématiques et computationnelles qui se contredisent mutuellement et reposent sur des normes de validité différentes. Néanmoins, comme on l'a vu pour les techniques et outils de simulation numérique dans les sciences de la nature, les techniques et modèles développées au sein de théories distinctes peuvent, du fait de leur caractère opératoire, circuler quel que soit la théorie

[32] https://fr.wikipedia.org/w/index.php?title=Apprentissage_automatique&oldid=134748222#Algorithmes_utilis.C3.A9s (consulté le 16 novembre 2017)
[33] http://bulletin.imstat.org/2014/09/data-science-how-is-it-different-to-statistics%E2%80%89/ (consulté le 16 novembre 2017)
[34] https://blog.mixpanel.com/2016/03/30/this-is-the-difference-between-statistics-and-data-science/ (consulté le 16 novembre 2017)

sous-jacente. Les modèles sont mobilisés de manière a-théorique, pour leur seule effectivité calculatoire, et non pour l'objet qu'ils ont pu représenter initialement.

La contribution de la statistique aux sciences des données, par ses méthodes et techniques, n'est de ce fait pas un phénomène isolé : plusieurs lignées historiques mobilisant des traitements computationnels ou pré-computationnels de données ont été agrégées dans le creuset des sciences des données. Si celle des statistiques se fait au prix d'une décapitation épistémologique, qui dépossède la statistique de ses normes théoriques, la contribution du champ de l'intelligence artificielle se fait semble-t-il d'une manière moins conflictuelle, due à son histoire propre et notamment à la façon dont ce projet a de lui-même révisé ses ambitions au cours du temps.

7. La contribution de l'intelligence artificielle

Le projet initial de l'intelligence artificielle peut être résumé comme un projet de conception d'artefacts visant à une meilleure compréhension de l'intelligence humaine, à la fois au niveau biologique, de la physicalité du cerveau, et au niveau cognitif, du fonctionnement de l'esprit. Comme l'ont souligné notamment Bruno Bachimont (1996) et Jean Lassègue (1996), ce n'est originellement pas un projet technologique de développement d'instruments de calcul au service des sciences cognitives mais un projet *de* sciences cognitives qui vise à une meilleure compréhension d'objets comme l'esprit ou l'intelligence à travers la construction d'outil de modélisation et de simulation de ces objets. Il faut l'entendre comme une science de l'esprit artificiel. À ce titre, elle produit et mobilise des thèses psychologiques ou neurologiques comme le computationnalisme, le connexionnisme, le cognitivisme et le behaviorisme. La conférence de Dartmouth de 1956, généralement considérée comme l'acte de naissance du projet de l'intelligence artificielle, se donne ainsi la finalité suivante :

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.

Il s'agit d'un projet d'investigation instrumenté par la machine, dont la mobilisation impose de formaliser les objets étudiés. Les grandes thématiques de l'intelligence artificielle y figurent déjà : le raisonnement, l'apprentissage du langage, la construction de réseaux de neurones artificiels, la relation entre machine et programmation, et la relation entre machine et esprit. C'est aussi d'emblée un projet d'artéfacture, qui vise à fabriquer des machines intelligentes.

La conférence de Dartmouth symbolise la rencontre entre deux chantiers de recherche actifs depuis les années 1940 : la cybernétique et les systèmes formels. Avec McCulloch et Pitts d'une part, sur lesquels nous reviendrons, mais aussi Wiener, la cybernétique avance alors déjà sur les questions de la relation entre pensée et calcul, et de la modélisation et construction de systèmes artificiels, suivant des approches plus ou moins tournées vers la modélisation ou vers l'ingénierie. De même que le concept de contrôle s'appliquait au vivant et à l'artificiel, celui d'intelligence s'appliquera à l'humain

comme à la machine. D'autre part, un chantier de recherche mathématique, animé notamment par Hilbert puis Gödel, aboutit aux travaux de Turing et une conception mécanique du raisonnement.

Néanmoins, l'intelligence artificielle est un projet vaste et ambitieux, qui agrège dès le départ des recherches diverses et des hypothèses de travail plus ou moins homogènes. Un même chercheur peut faire évoluer son approche et ses hypothèses. Ainsi, le même Turing, qui répond en mathématicien au problème de la décision de Hilbert en 1936 s'interroge, une quinzaine d'années plus tard, sur la nature de la pensée et sur les conséquences de l'hypothétique mise au point de machines pensantes. Avant même la conférence de Dartmouth, l'article *Computing Machinery and Intelligence* de 1950, publié dans la revue de psychologie et de philosophie *Mind*, pose un programme de recherche pour la science de l'esprit artificiel. Dans cet article, il commence par revenir sur l'ambition de créer « des machines qui pensent » en évacuant la pertinence de la question même de déterminer si une machine peut penser, et en la remplaçant par la question de la capacité d'une machine à imiter le comportement d'un humain. C'est le fameux « jeu de l'imitation », désormais appelé « test de Turing », alors posé comme une expérience de pensée pour reformuler la question de l'articulation entre pensée et machines.

Turing vise l'imitation des capacités cognitives humaines, excluant celle du corps humain incarnée ultérieurement par une partie de la recherche en robotique. S'il considère initialement comme machine tout artefact similaire à l'homme, y compris un homme qui serait bioingéniéré, en tous points semblable à un homme « né par les voies naturelles », sa réflexion s'oriente ensuite plus spécifiquement sur les calculateurs numériques *(digital computers)* ou, dirions-nous aujourd'hui, des ordinateurs. Si la reformulation des capacités de ces machines en capacité d'imitation plutôt qu'en capacité de pensée peut paraître à première vue comme une réduction de cette ambition, il s'agit au contraire d'une manière de contourner un certain nombre d'objections relatives à la pensée et la conscience des machines. Le projet de l'intelligence artificielle est ainsi redéfini comme un projet d'artéfacture non pas de machines pensantes, mais de machines capables d'imiter de manière suffisamment convaincante les capacités cognitives humaines. Formulé ainsi, ce projet fait l'objet d'une spéculation extensive de la part de Turing, qui va jusqu'à se demander comment ses machines et les humains pourraient se lier d'amitié, ou comment bien intégrer des « machines-enfants » à l'école sans que les autres élèves ne s'en moquent de façon excessive. Enfin, la conclusion de l'article propose un certain nombre d'options pour les travaux futur en intelligence artificielle :

> We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

> We can only see a short distance ahead, but we can see plenty there that needs to be done.

Turing ébauche un programme de recherche qui sera remarquablement suivi. Reconnaissance de la parole, reconnaissance de forme, traitement automatique des langues, agents conversationnels, systèmes experts, apprentissage automatique, sont autant de champs de recherche dont le succès est considéré comme nécessaire pour fabriquer des machines qui apprennent, perçoivent, parlent, comprennent et raisonnent.

Le Perceptron imaginé par Frank Rosenblatt en 1958 constitue l'un des premiers, voire le premier exemple de machines effectivement construites pour réaliser ce programme. Il reflète la dualité entre science et technique du projet initial de l'intelligence artificielle, dans la mesure où il s'agit à la fois d'une description théorique du cerveau comme système de stockage et de traitement d'information, et de la conception d'un modèle computationnel correspondant à cette définition :

> The theory to be presented here takes the empiricist, or "connectionist"' position with regard to [the] questions [that we must answer to understand the capability of higher organisms for perceptual recognition, generalization, recall, and thinking]. The theory has been developed for a hypothetical nervous system, or machine, called a *perceptron*. The perceptron is designed to illustrate some of the fundamental properties of intelligent systems in general, without becoming too deeply enmeshed in the special, and frequently unknown, conditions which hold for particular biological organisms. The analogy between the perceptron and biological systems should be readily apparent to the reader.

Comme on peut l'attendre d'un travail qui s'inscrit dans le champ des sciences cognitives, publié dans une revue de psychologie, l'article de Rosenblatt examine dans un premier temps les différentes théories et positions existant à son époque sur le traitement de l'information dans le cerveau, et revendique son appartenance à l'école connexionniste. Il faut d'ailleurs noter que le fait même de conceptualiser le système nerveux comme un système de traitement d'information est lui aussi un postulat implicite, qu'on pourrait discuter ou formuler en d'autres termes. L'article propose ensuite un modèle fonctionnaliste du cerveau, dégagé de son substrat biologique, mais imitant sa structure, et dont l'ambition est de reproduire et représenter son fonctionnement : Rosenblatt conclut en espérant explicitement que son modèle aide à mieux comprendre « les systèmes de traitement d'information, humains aussi bien qu'artificiels ». Il s'intéresse plus spécifiquement à la perception, la reconnaissance visuelle et l'apprentissage. Il ne s'agit pas d'un modèle purement conceptuel, mais bien d'un modèle computationnel et d'une procédure formelle, autrement dit d'un algorithme, avec son implémentation logicielle et matérielle, dont le Perceptron Mark I (voir **Figure 8**) développé dans les années 1950 au Cornell Aeronautical Laboratory.

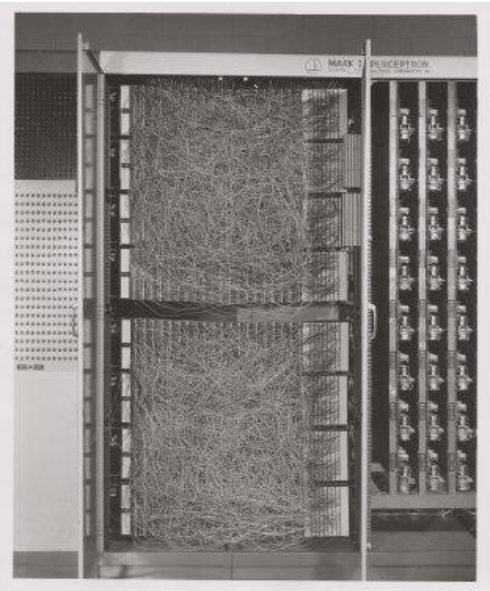

Figure 8*. Le Perceptron Mark I au Cornell Aeronautical Laboratory*

Ce modèle est conçu pour être dynamique : suivant l'idée que le système nerveux se reconfigure en fonction des stimuli visuels qu'il reçoit, un perceptron se reconfigure en fonction des informations qui lui sont envoyées. On peut ainsi considérer le perceptron comme l'algorithme fondateur de l'apprentissage automatique *(machine learning)*, c'est-à-dire d'une procédure formelle capable de modifier la structure de son modèle en fonction de l'information qui lui est envoyée, ou comme on le dira aujourd'hui, d'apprendre à partir de données. Cette capacité d'apprentissage est la contribution majeure du perceptron par rapport aux réseaux de neurones artificiels imaginés par McCulloch et Pitts (1943), qui propose un modèle du neurone plus sophistiqué d'un point de vue logico-formel, mais incapable d'imiter la plasticité du cerveau et d'apprendre à partir de l'information reçue.

S'il l'on compare l'ambition de Rosenblatt à celle de Turing, on notera qu'il ne s'agit plus de construire une machine qui imiterait le comportement d'un humain de manière suffisante convaincante pour un évaluateur humain, mais de fabriquer quelque chose qui ressemble au cerveau d'un point de vue fonctionnel, avec la nécessité de proposer un modèle dudit cerveau. Dans les termes des sciences cognitives, Rosenblatt est cognitiviste et représentationaliste là où Turing est behavioriste, incarnant ainsi plusieurs conceptions du cerveau de leur époque.

Sévèrement critiqué par Marvin Minsky et Seymour Papert en 1969 dans leur livre *Perceptrons* pour son incapacité à apprendre effectivement des fonctions logiques élémentaires, le perceptron est cependant le premier avatar d'une lignée technique de réseaux de neurones artificiels apprenants dans laquelle s'inscrivent d'abord des techniques supervisées (comme les réseaux de neurones récurrents introduits par Stephen Grossberg (1988) et les réseaux de neurones convolutionnels introduits par Kunihiko Fukushima (1980)) puis des techniques non supervisées à partir de la toute fin des années 1980 (LeCun, 2014). Cette lignée a longtemps été en disgrâce à la fois pour la faiblesse de ses résultats, l'insuffisance de la puissance de calcul des ordinateurs, et la domination d'une part du courant

symboliste et de l'ingénierie des connaissances en IA, et de techniques d'apprentissage automatique plus simples et efficaces comme les SVM et les random forests. Cette disgrâce conduit notamment à une réduction de la portée théorique des réseaux de neurones artificiels, qui ne date donc pas de l'avènement des sciences des données (sauf à réécrire rétrospectivement l'histoire de ces domaines). Au début des années 1980, chez Kunihiko Fukushima (1980) par exemple, le projet de l'intelligence artificielle et de compréhension du cerveau est encore vivant :

> If we could make a neural network model which has the same capability for pattern recognition as a human being, it would give us a powerful clue to the understanding of the neural mechanism in the brain.

Dès la moitié des années 1980, le ton a changé. Les notions de cerveau et de systèmes nerveux disparaissent, et les objectifs de recherche relèvent désormais du perfectionnement logico-mathématique et computationnel. Dans un article de 1989 jugé influent par Yann LeCun, Hornik, Stinchcombe, & White (1989) décrivent ainsi leur ambition dans les termes suivants :

> This paper rigorously establishes that standard multilayer feedforward networks with as few as one hidden layer using arbitrary squashing functions are capable of approximating any Borel measurable function from one finite dimensional space to another to any desired degree of accuracy, provided sufficiently many hidden units are available.

De même, le travail de David Rumelhart, Geoffrey Hinton et Ronald Williams (1985), qui ont lu et citent le travail de Minsky et Papert, relève du perfectionnement des techniques :

> The major theoretical contribution of the work is the procedure called error propagation, whereby the gradient can be determined by individual units of the network based only on locally available information.

Cette lignée est revenue en force à la suite de l'apparition du phénomène *big data*, à la faveur d'une configuration qui bénéficie de ces perfectionnement des techniques, d'une puissance de calcul et un volume de données inédits.

La différence majeure entre les réseaux de neurones artificiels utilisés aujourd'hui dans les sciences des données et ceux de Rosenblatt ou de McCulloch et Pitts, est, comme on le voit dès la fin des années 1980, qu'ils sont désormais amputés de leur ambition théorique initiale pour les chercheurs et praticiens de ces techniques : on ne retrouve pas la trace d'une volonté d'artéfacture d'une intelligence artificielle générale, ni de recherche explicative en sciences cognitives[35]. Il ne s'agit plus de représenter mais tout au mieux de reproduire, d'imiter des capacités cognitives comme la vision, la parole ou le langage. C'est là la mutation essentielle qui permet de comprendre la différence entre

[35] L'imaginaire associé à l'intelligence artificielle est cependant lui bien vivant, avec son lot d'enthousiastes et d'inquiets. Dans les mois où cette thèse s'est écrite, on a vu les discours médiatiques et commerciaux passer de « le big data a permis de X » à « un algorithme a permis de X » puis à « une intelligence artificielle a permis de X », pour décrire un procédé identique. Que le procédé en question repose au moins sur de l'apprentissage automatique ou sur un ensemble de règles formalisées manuellement n'intervient que rarement dans le choix du vocabulaire en question.

l'intelligence artificielle, projet à la fois scientifique, technologique et artéfactuel, et l'apprentissage automatique comme champ de recherche et d'application technologique. Si l'une peut être vue comme une technoscience productrice d'artefacts à visée explicative, l'autre perd cette visée explicative. Comme pouvait le laisser prévoir le programme de Turing dans son article de 1950, des champs plus spécialisés, comme la robotique, le traitement automatique du langage, l'ingénierie des connaissances ou la vision artificielle, se sont également développés. Passer d'un programme général à un ensemble de champs spécialisés, c'est au moins remettre à plus tard le projet d'une intelligence artificielle forte, semblable à l'intelligence humaine, qui serait constituée de l'assemblage des résultats de ces différents champs spécialisés. En pratique, ces champs sont également passés de domaine de recherche scientifique à visée explicative, à domaine de recherche technologique à visée instrumentale. Les questions théoriques, voire philosophiques, sont généralement évacuées, à moins qu'un problème pratique ne puisse être résolu sans revenir à un niveau plus conceptuel. Avec l'efficacité des techniques d'apprentissage automatique dans ces domaines, les applications industrielles deviennent à portée de main, ce qui favorise également un développement instrumental appliqué (produire et perfectionner des objets techniques) au détriment d'un développement fondamental.

Plus prosaïquement, les réseaux de neurones artificiels sont utilisés essentiellement comme n'importe quel algorithme de classification automatique, au même titre que les SVM ou les random forests. Les techniques de fouille de données issues du développement des entrepôts de données, les réseaux de neurones inspirés par le programme initial de l'intelligence artificielle, et les autres techniques d'apprentissage automatique, ont eu des développement historiques distincts, mais ne présentent pas de distinction théorique ou fonctionnelle forte : il n'y a guère de bonne justification, à proposer une classification des techniques mobilisées par les sciences des données qui suive ces distinctions historiques, d'autant que ces lignées se sont maintes fois croisées avant l'apparition des sciences des données. Le travail de Gregory Piatetsky-Shapiro et ses collègues (1996) sur la découverte de connaissances dans les bases de données (KDD), et que nous avons déjà évoqué plus haut, a d'ailleurs été publié dans *AI Magazine*, la revue de l'American Association for Artificial Intelligence. Ils y soulignent que le terme s'est popularisé depuis 1989 dans les champs de l'intelligence artificielle et de l'apprentissage automatique ; il y décrit également la KDD comme un champ émergeant de la rencontre de champs existants, dont l'IA :

> KDD has evolved, and continues to evolve, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and high-performance computing.

Dès les années 1990, la distinction entre intelligence artificielle (entendue au sens d'ensemble de techniques d'apprentissage automatique) et fouille de données (ou KDD) est devenue floue. Le projet de science de l'esprit artificiel n'est plus guère à l'œuvre dans ces développements, même si certains chercheurs ou ingénieurs restent des vecteurs de croyances computationnalistes. Le développement

des entrepôts de données coïncide historiquement avec l'apparition de techniques d'apprentissage automatique non supervisées (dont des techniques à base de réseaux de neurones artificiels) également utilisées pour la fouille de ces entrepôts de données. L'apparition de nouvelles données dans les entreprises, le développement d'un creuset de techniques, constitue une première conjonction entre de nouvelles données, un ensemble de techniques et une amélioration de la puissance de calcul, qu'on pourrait voir comme un premier âge des *big data* et des sciences des données. Le développement de la fouille de données intègre en tous cas les différentes caractéristique des *data sciences* comme art du calcul : un savoir-faire fondé sur une recherche autonome de techniques computationnelles, une prééminence des modèles computationnels sur les modèles conceptuels, l'insertion dans un paradigme exploratoire, et une circulation des techniques indépendamment du projet épistémique rendue possible par ces différentes caractéristiques.

Conclusion

Le calcul ou analyse computationnelle est une composante nécessaire du paradigme méthodologique permettant de dégager une intelligibilité des *big data*. Il fait partie des sciences computationnelles de la culture mais n'en constitue pas la totalité : il est entouré, en amont, par un cadre théorique et technique permettant de conférer une valeur épistémique à la donnée et de constituer une unité cohérence d'information (un corpus), et en aval, par le travail d'interprétation et de recontextualisation que nous aborderons dans les chapitres suivants.

Sur le plan épistémologique, la mobilisation du calcul engendre une double restriction du connaissable au calculable :

- la donnée doit devenir calculable, en terme de format et de support, pour faire intervenir les sciences des données ;
- le calculable et le calculé déterminent le champ du connaissable et constituent un sous-ensemble de celui-ci par l'intermédiaire de la mathématisation des problèmes épistémiques.

En d'autres termes, il y a une double détermination de la connaissance et de la donnée par le calcul, qui définit les conditions de possibilités de la production de l'une à partir de l'autre. À cette double détermination théorique s'ajoute une restriction pratique de ce qui est connaissable, à travers une culture épistémique spécifique, la pensée computationnelle, qui oriente la production de connaissances vers les objets susceptibles d'être soumis à une modélisation et un calcul.

Cette double détermination est commune à toutes les mobilisations du calcul dans le traitement de données en vue de produire des connaissances, qu'il s'agisse des sciences de la nature, des disciplines à terrain ou des sciences des données. Dans les sciences de la culture, la mobilisation du calcul procède principalement d'un travail de modélisation, à partir duquel le calcul occupe un statut proche de l'observation ou de l'expérience. La simulation numérique, en particulier, permet de produire des

données qui se substituent à l'expérience, et remplacent les données expérimentales obtenues *via* les mesures d'instruments scientifiques par des données simulées. Le calcul s'intègre alors dans une continuité épistémique qui va de la théorie à l'expérimentation en passant par le modèle. À l'inverse, les sciences des données ont un statut de savoir-faire sans cadre théorique, qui n'assigne pas de signification spécifique à la donnée. Elles interviennent au niveau du modèle computationnel et non aux niveaux supérieurs. Elles prennent la donnée pour elle-même, celle-ci constituant le territoire à explorer au moyen de techniques computationnelles dont les origines théoriques et institutionnelles sont multiples et sans conséquence sur les pratiques et les résultats. À ce titre, elles constituent un creuset de techniques qui joue le rôle d'une boîte à outil exploratoire au service de projets épistémiques variés. Ces projets peuvent être définis préalablement ou émerger de l'exploration, et ne recevront de signification définitive qu'ultérieurement.

Les sciences des données ne fournissent pas en effet de normes ou de cadre conceptuel qui permettrait d'interpréter le résultat du calcul comme une certaine représentation du réel. Cela signifie que la valeur épistémique de la donnée, obtenue préalablement au calcul, est suspendue par celui-ci, et que la signification des résultats, tout en étant cohérents par rapport à cette valeur épistémique, doit être reconstruite en attribuant une signification aux transformations computationnelles de la donnée. En ces termes, le projet épistémique n'a pas tant pour fonction de fournir un cadre *a priori* au traitement des données qu'un ensemble d'hypothèses de travail qui émergent et évoluent au fil de l'exploration. Il est le fil conducteur qui permet d'expliciter une logique de constitution de corpus, d'orienter la préparation, le formatage et l'exploration des données, et enfin de formuler une signification communicable à travers un savoir-faire interprétatif.

C'est en effet au praticien qu'il revient de découvrir le sens des données à partir de son projet épistémique en mobilisant itérativement l'imagination et le calcul suivant un paradigme exploratoire. La valeur épistémique des données est un possible actualisé par ces deux opérations de l'esprit outillé, mis en suspens par le calcul. Comme nous allons maintenant le voir avec un cas concret d'exploitation des *big data*, cet art de l'interprétation est présent à toutes les étapes de constitution et de restitution du sens des données, que cette restitution soit narrative, visuelle ou logicielle. Quoique cet art relève, enfin, largement des méthodes des sciences de la culture, c'est dans un tout autre cadre institutionnel, celui de l'édition logicielle, que nous allons le retrouver.

Troisième partie.

Interpréter, visualiser, décider

Chapitre VI.
Un cas concret d'exploitation de données massives

Nous avons jusqu'à présent mis en évidence les mythologies et pratiques associées aux données massives ainsi qu'à leur traitement computationnel. La mythologie des *big data* s'oppose aux modalités concrètes d'exploitation des traces numériques issues du web, tandis que les *digital humanities* représentent de manière putative une configuration réalisée en pratique par les data sciences, elles-mêmes chargées de la mythologie de l'intelligence artificielle. Le statut de la donnée et des techniques de traitements constituent à la fois des nœuds de questionnement épistémologique, et des étapes dans les processus de production de connaissances à partir de traces numériques. Le statut de la donnée se pose avec la constitution du corpus, tandis que ce que les *data sciences* font aux données se matérialise lors de leur traitement effectif. En explicitant ces deux composantes que sont la donnée et le calcul, nous avons également mis en évidence la nécessité simultanée d'une troisième composante et d'une troisième étape pour apporter une intelligibilité aux données : la question de l'interprétation des résultats du calcul, et le rôle de l'herméneutique de la production de connaissances valides.

Par ailleurs, la façon dont nous avons interrogé la donnée et le calcul est restée jusqu'à maintenant plutôt générale, avec une ambition de décrire les caractéristiques communes à tous les traitements computationnels de traces numériques. Si l'analyse des conditions générales d'émergence de connaissances dans ce contexte a constitué une partie conséquente du travail de cette thèse, l'examen d'un cas plus précis en a également été une composante significative, dont nous avons retardé la présentation jusqu'à maintenant. Afin de donner du poids à cette généralité, de montrer la façon dont elle se ramifie en s'ancrant dans le réel, nous allons maintenant présenter ce cas plus précis en décrivant comment les problèmes de constitution de corpus et de nettoyage de données, les modalités de traitement et ce qu'elles font à la donnée, se matérialisent concrètement dans une situation donnée. Nous pourrons ainsi confirmer et nuancer ce que nous avons développé jusqu'à présent, tout en montrant que chaque situation présente également des problèmes épistémologiques spécifiques, qui ne peuvent être considérés comme des propriétés universelles du traitement de traces numériques, mais qui illustrent en revanche la densité des ramifications méthodologiques qu'il implique.

Ce cas plus précis s'appuie sur une expérience de terrain – dont nous décrivons ci-dessous les modalités – qui a permis d'observer et de participer plusieurs fois à un processus de traitement de traces numériques. Tout en étant plus précis, il conserve également une certaine généralité par rapport à un processus unique qui aurait été observé comme un événement singulier ; il constitue en effet une synthèse des répétitions du processus. Il illustre de plus plusieurs points que nous avons développés jusqu'à maintenant : il s'agit, bien sûr, de traces numériques relatives à des comportements humains, et qui font l'objet d'un traitement computationnel avec des techniques inspirées de l'intelligence artificielle, de l'informatique décisionnelle et de la statistique. Ces traces numériques ne sont pas exploitées par des chercheurs en sciences de la culture, mais pas une entreprise éditrice de logiciel pour le compte de clients. Cette entreprise possède d'une part la culture technique pour manipuler de telles données, et d'autre part réalise ces opérations dans une finalité non scientifique.

Ce cas possède cependant des caractéristiques singulières par rapport à la situation archétypique que nous avons dessinée jusqu'à présent. En effet, les données analysées sont de nature textuelle, ce qui mobilise des techniques et problématiques de format et de représentation de données supplémentaires par rapport aux données dites structurées. De plus, leur analyse est réalisée de manière distribuée entre plusieurs agents épistémiques qui occupent des fonctions différentes et se comprennent donc, comme on le verra, dans le cadre d'une épistémologie sociale.

1. Présentation de l'étude de cas

Cette étude de cas repose sur une expérience prolongée sur plusieurs années au sein d'une organisation qui traite massivement des données, l'éditeur de logiciel Proxem. Dans sa communication, Proxem se présente comme un acteur des *big data* et constitue ainsi un exemple de relai de cette mythologie ; en termes d'activité, Proxem est une startup technologique spécialisée dans le traitement automatique du langage, une branche de l'intelligence artificielle consacrée à l'analyse de contenus textuels numériques, et qui fait partie des techniques mobilisées dans les data sciences. En termes de mythologie comme en termes de pratique, Proxem est un acteur du traitement computationnel de traces numériques.

Je fais partie de la société Proxem depuis 2012 ; mes différentes fonctions m'ont amenées à analyser, comprendre, concevoir, des traitements de données textuelles. La nature des données, la complexité des traitements, les difficultés à dégager des résultats qui avaient du sens, à évaluer leur fiabilité, et à relier ces résultats à des besoins clients, ont déclenché une réflexion qui a été le point de départ de cette thèse. Cette réflexion a été individuelle, avec le présent travail de recherche, mais aussi collective, *via* les échanges, les discussions, les expérimentations de mes collègues. J'ai ainsi pu observer leurs pratiques, leurs modes de travail, leur façon d'aborder un problème, mais aussi

échanger et contribuer de manière active à leur activité, sans rester dans une posture d'observation pure.

Le cas présenté relève ainsi d'une démarche ethnographique, dans la mesure où il puise sa légitimité dans son ancrage au terrain, mais aussi d'une logique d'intervention forte dans les phénomènes observés, qui déborde du cadre de l'observation participante. Il ne s'agissait pas d'être *data scientist*, et de décrire de l'intérieur les pratiques de ce type d'acteur, mais de faire partie d'un travail collectif sur plusieurs années mobilisant des data scientists, des ingénieurs logiciels, des ingénieurs d'affaires, des communicants, des linguistes, des managers, des dirigeants, etc. Ce n'est donc pas une ethnographie du *data scientist*, qui permettrait de dégager l'essence de cette figure, mais une recherche dans une structure qui intervient sur des problématiques de data science, dont on cherche à montrer la complexité, et notamment la diversité d'acteurs, de fonctions et de postures épistémiques pour aborder collectivement cette problématique. De ce fait, on peut voir cette intervention comme une recherche-action dans une structure pertinente par rapport à l'objet de la thèse, qui constitue à la fois une recherche sur les traitements de traces numériques visant à produire une compréhension nouvelle de cet objet, et un travail productif qui intègre cette recherche dans des projets client et la conception du logiciel ; action et production se nourrissent mutuellement et se transforment à mesure qu'elles interviennent l'une dans l'autre. Chez Lewin (1952) en effet, la compréhension passe par la modification : la description ou observation des pratiques est considérée comme insuffisante, tandis que l'action permet vraiment de comprendre l'objet modifié.

La notion de recherche-action, qui présente de nombreuses ramifications (Jouison-Laffitte 2009), décrit bien, dans sa forme classique, l'accord mutuel conclu avec Proxem relatif à mon travail de thèse et son articulation avec mes fonctions dans l'entreprise. Tandis que l'ancrage chez Proxem donnait à ma thèse une substance dans laquelle puiser pour problématiser son objet, ce travail de recherche devait fournir à l'entreprise une contribution méthodologique, permettant d'améliorer les offres de services et produits logiciels, et idéalement constituer, sous la forme d'une méthode propriétaire, un produit commercialisable en soi. L'accord conclu m'a donné une position privilégiée pour observer les pratiques de l'entreprise et illustrer les processus des sciences des données. La connaissance détaillée de ces pratiques m'a permis d'adapter ma contribution aux spécificités de l'entreprise, et de proposer une intervention adaptée à ses problématiques. Il y a ainsi eu, comme le préconise Robinson (1993), convergence entre les intérêts théoriques de la thèse, et les attentes de l'entreprise quant à celle-ci.

Toutes les références à la conception logicielle, au traitement du langage, à la gestion de projets client, etc., qui vont suivre dans ce chapitre et les suivants, sont nourries par cette expérience de terrain qui prend la forme d'observations, d'entretiens informels, de formations internes, d'échanges, de contribution, de pratique, de discussion, etc. Jusqu'en 2015, toute l'équipe était réunie dans un même bureau, ce qui m'a permis d'être physiquement témoin de quasiment tous les échanges entre membres

de l'équipe, par immersion physique. Je les ai vus et entendus travailler, j'ai pu voir leurs écrans, les outils qu'ils utilisent, etc. et ainsi observer ce que c'est concrètement que de travailler chez un éditeur de logiciel. Ma compréhension de ce qu'est par exemple une mise en production dans l'industrie logicielle provient à la fois des explications de mes collègues, de leurs conversations entre eux avant, pendant et après l'opération, de mon accès aux notifications dans les outils internes de *monitoring* des différentes étapes de l'opération, de mon accès au logiciel avant et après, et de la répétition de ces points jusqu'à acquérir une certaine vision de la variabilité de l'opération, de la façon dont elle peut bien ou mal se dérouler, de ce qu'elle implique, etc. À cela s'ajoute l'accès à d'autres informations en dehors de l'entreprise, *via* des événements professionnels, des rendez-vous avec des clients et prospects, et des échanges informels et amicaux avec d'autres acteurs du secteur, qui ont confirmé les témoignages de mes collègues.

Le fil directeur de l'étude de cas que nous allons présenter est le processus d'analyse d'un certain type de données. Comme nous l'avons évoqué, il ne s'agit pas de la présentation d'un projet unique dans sa singularité, mais déjà d'une synthèse de différents projets réalisés pour des clients de Proxem. Cette synthèse ne provient pas de l'observation seule du déroulement des projets, mais détaille le fonctionnement des outils utilisés, les motivations des méthodes et façons de faire, les explications des différents participants, de manière à décrire mais aussi à expliquer ce processus. De plus, cette synthèse se nourrit de mon propre travail de recherche, des connaissances universitaires acquises, en parallèle de mes fonctions, sur la linguistique, le traitement automatique des langues, la statistique, les méthodes quantitatives et qualitatives en sciences sociales, la textométrie, etc., auxquelles je fais référence dans cette synthèse lorsque cela permet d'apporter un éclairage supplémentaire. La mobilisation de ces connaissances universitaires a en effet permis de mieux comprendre ce processus, de l'inscrire dans des lignées méthodologiques qui étaient connues ou non de mes collègues, et qui ont facilité ma contribution au processus.

Le processus en question inclut des étapes :

- de constitution et de nettoyage des données ;
- d'analyse de données textuelles ;
- d'interprétation, de visualisation et de restitution.

Si les deux premières, abordées dans les chapitres qui précèdent, étaient en partie stabilisées lors de mon arrivée dans l'entreprise, les étapes finales étaient encore émergentes. Ma recherche-action a ainsi apporté des clés théoriques, méthodologiques et pratiques pour les développer, en construisant un savoir-faire interprétatif nourri de mes connaissances universitaires sur l'analyse de données dans les sciences de la culture. Ces connaissances m'ont permis de passer d'une contribution empirique et pratique au processus d'analyse de données, à une contribution plus formelle permettant d'expliciter ses modalités méthodologiques et théoriques et d'être formulée et transmise dans l'entreprise. Du fait

que ma contribution intervenait principalement sur les phases finales du processus, il était impératif de bien comprendre les étapes précédentes pour leur donner sens. En d'autres termes, là où mes collègues intervenaient principalement sur la constitution de la donnée et le calcul, étapes abordées dans les chapitres précédents, j'ai contribué à l'interprétation des résultats du calcul, qui sera analysée plus spécifiquement par la suite.

Dans ce qui suit, la présentation du processus n'est donc pas tant la description brute d'une série d'événements, qu'une reconstruction reconceptualisée à partir de plusieurs exemples, de la procédure type de traitement de données dans l'organisation. La reconstruction repose à la fois sur des observations empiriques, et sur des apports théoriques, des rapprochements avec d'autres pratiques épistémiques, dans un but de clarification de la procédure. Elle mêle ainsi d'un côté des observations de pratiques, des descriptions de celles-ci par leurs praticien-ne-s, et de l'autre une contribution épistémologique qui formalise ces pratiques, les relie, les inscrit dans des traditions de recherche, pour en expliciter le sens.

Proxem est un éditeur de logiciel fondé en 2007 par François-Régis Chaumartin, entrepreneur, ingénieur en informatique et docteur en linguistique (depuis 2012, voir Chaumartin 2012), qui commercialise des solutions d'analyse sémantique fondées sur des technologies de traitement automatique du langage. Proxem propose plusieurs domaines d'application de ses technologies et services, appelés « expertises ». Les trois principales expertises de l'entreprise sont définies de la façon suivante :

- l'**expérience client**, où Proxem aide les entreprises à analyser les discours des consommateurs et de leurs clients ;
- les **ressources humaines**, où sont analysés les documents associés aux salariés et candidats d'une entreprise (offres d'emploi, CV, entretiens annuels, etc.) ;
- la **market intelligence**, en vue de quoi est analysée la présence sur le web des acteurs économiques, publics et parties prenantes d'un marché.

L'offre commerciale de l'entreprise est par ailleurs structurée en des produits de différente nature :

- un **logiciel** d' « exploration et de visualisation de données textuelles » disponible par abonnement, qui permet à ses utilisateurs, depuis mars 2017, de :
  - collecter des contenus textuels sur le web ou d'en charger dans le logiciel à partir d'un tableau Excel ;
  - extraire de l'information de ces contenus et la structurer en fonction d'un « thésaurus », aussi appelé « plan de classement » ou encore « dictionnaire de concepts » ;
  - quantifier l'information et la visualiser de manière interactive ;

    - partager les résultats sous formes d'exports, de tableaux de bord ou de rapports personnalisés ;
  - un ensemble de **services** de formation, d'accompagnement, de paramétrage du logiciel et de production d'analyses sur mesure.

Depuis mon entrée dans l'entreprise en février 2012 (soit plus d'un an avant le démarrage du travail de recherche matérialisé par le présent manuscrit), mes fonctions ont été multiples et ont évolué. Je suis notamment (mais pas exclusivement) intervenue dans :

- la structuration et la promotion des offres commerciales ;
- l'élaboration de rapports d'étude et la conception d'une méthode de production de ces rapports ;
- la conception et l'analyse des usages du logiciel.

2. Expérience client et codification de verbatims

Dans un but de lisibilité et de synthèse, nous ne présenterons que le processus lié à l'expertise appelée « expérience client » : présenter le déroulé complet d'un projet type de cette nature fournit en effet une meilleure illustration des usages concrets des sciences des données chez Proxem que de tenter d'agréger des pratiques relevant des différentes expertises, et chacune régie par des normes spécifiques. En particulier, l'expertise « market intelligence » repose sur la constitution de corpus de données à partir du web, un exercice encore faiblement stabilisé dans l'entreprise (donc difficile à conceptualiser) et qui soulève des enjeux techniques et épistémologiques que nous n'allons pas développer ici pour nous concentrer sur le rôle de l'instrumentation technique et de l'interprétation dans le processus d'analyse de données.

L'expérience client, précédemment appelée connaissance client, est la plus ancienne expertise de l'entreprise, la plus stabilisée, et celle dont ont, en substance, découlé les autres ; nous la présentons telle qu'elle est proposée par Proxem en 2017. Parmi les fonctions que j'exerce ou ai exercé chez Proxem, un rôle de chargée d'études, puis responsable d'études, m'a permis d'être en contact direct avec les destinataires des projets qui s'inscrivent dans cette expertise, et ainsi de comprendre leur métier, leurs problématiques, leurs attentes, et leurs critères d'acceptabilité du travail réalisé par Proxem.

Avant d'être un domaine d'application des compétences technologiques et méthodologiques de Proxem, l'expérience client est en effet une notion issue du marketing, qui désigne les rapports entre une entreprise et ses clients, et les conceptualise du point de vue de l'expérience subjective que les clients ont de l'entreprise à travers ses différentes composantes (ses produits, ses points de vente, sa communication, son identité, sa réputation). Proxem a développé sa connaissance de cette notion au contact des clients et prospects qui exprimaient un besoin quant à ce sujet, afin de promouvoir ses

technologies dans des termes intelligibles pour ces acteurs. C'est une notion qui recoupe partiellement des problématiques d'expérience utilisateur (où l'on se concentre sur l'utilisation spécifique d'un produit), d'études marketing (qui visent à analyser le marché, c'est-à-dire le point de rencontre entre l'offre et la demande) ou encore d'études de communication (qui analysent notamment la perception de l'entreprise et de la marque par les consommateurs). L'expérience client est une problématique généralement gérée au sein d'une direction marketing ou études, par des postes créés spécifiquement pour ce besoin tels que des « responsables de la relation client », ou des postes préexistants à la conceptualisation de cette problématique, comme les responsables de centre d'appel ou les directeurs de divisions régionales.

Les principales missions des représentants de la fonction « expérience client » sont, dans leurs propres termes, les deux suivantes[36] :

- **recueillir de l'information** sur l'expérience des clients et leur satisfaction, c'est-à-dire développer de la « connaissance client » et suivre ses évolutions ;
- **répondre aux attentes des clients** c'est-à-dire améliorer leur satisfaction, mettre en place des actions empêchant les expériences négatives de se reproduire ou diminuer leur fréquence, recontacter individuellement les clients stratégiques, mettre en place et administrer un service client, téléphonique, physique ou numérique.

Dans le cadre de leurs fonctions, les responsables de l'expérience client doivent donc brasser beaucoup d'information, souvent de nature textuelle, émise par leurs propres clients. Dans ce but, ils peuvent se doter de prestataires externes (entreprises de sondage, cabinets de conseils, éditeurs de logiciel) ou développer un dispositif dédié en interne. Il peut s'agir d'enquêtes ponctuelles ou de baromètres de satisfaction, comme d'enquêtes dites « à chaud », à travers lesquelles le client est interrogé immédiatement après une action telle qu'un achat, un abonnement, ou un passage en magasin. Par exemple, s'il achète un produit sur un site marchand, l'entreprise responsable du site envoie un e-mail après réception de sa commande pour lui demander s'il est satisfait de son achat et s'il a des remarques. D'autres clients de Proxem sondent leurs clients immédiatement en point de vente : une chaîne de restaurants a ainsi doté ses employés de tablettes tactiles qui permettent de prendre des notes en interrogeant le client après un repas.

Ces dispositifs d'*écoute client* sont plus ou moins formalisés ; par exemple, de nombreuses entreprises ont adopté la méthodologie Net Promoter Score (NPS) déposée par le cabinet de conseil en management Bain & Company, et qui a par ailleurs fait l'objet de plusieurs livres (Brooks et Owen 2008; Reichfeld et Markey 2011). Dans le cadre de cette méthode, le client doit noter son expérience de 0 à 10, et justifier sa note en quelques mots. À partir de l'agrégation de ces notes, la méthodologie

[36] Fin 2017, une mission supplémentaire prend de l'ampleur, celle d'optimiser voire d'automatiser les processus de traitement des demandes *via* des technologies de type « assistant de réponse » ou « agent conversationnel » (*chatbot*).

permet de calculer un score (**Figure 9**) qui évolue dans le temps, ainsi qu'un certain nombre d'indicateurs matérialisés dans des tableaux de bord. À ce titre, c'est comme on le verra au chapitre IX, un dispositif de *benchmarking* (Bruno et Didier 2013) qui permet de scorifier et mesurer la performance de l'entreprise, suivre son évolution, et enfin comparer entre eux les acteurs, les points de vente, les divisions régionales, etc.

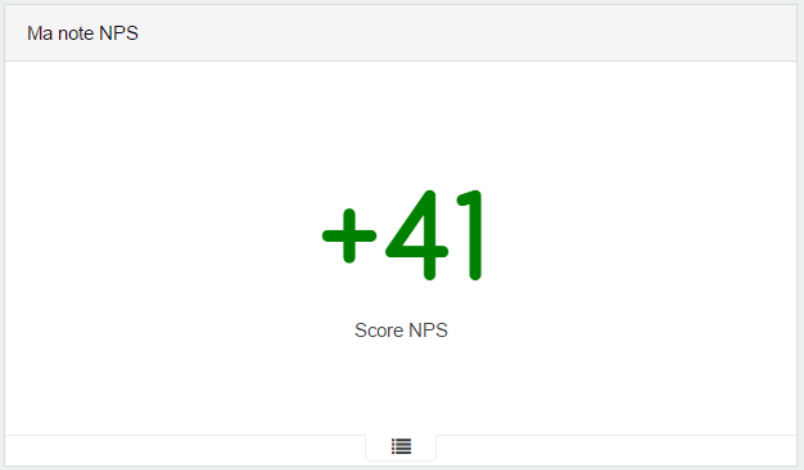

Figure 9. *Exemple de visualisation d'un score NPS dans un tableau de bord. Le score NPS est une mesure quantitative de la satisfaction client exprimée sur une échelle de -100 à +100. Il est calculé à partir de la note moyenne des clients satisfaits moins celle des clients insatisfaits.*

Les dispositifs de type NPS produisent beaucoup de réponses sous forme de texte libre, en raison, en l'occurrence, de la question ouverte qui permet de justifier sa note. Dans d'autres dispositifs, la question ouverte peut être plus ou moins facultative, et il peut y en avoir plusieurs : une question peut être associée à chaque produit, ou servir à exprimer la satisfaction du client en deux temps :

- « qu'est-ce qui vous a plu ? », « qu'avez-vous apprécié ? » ;
- « qu'est-ce qui vous a déplu ?, « qu'est-ce qui pourrait être amélioré ? »

Le rôle de Proxem dans les démarches d'expérience client est de procéder à la *codification* des réponses à ces questions ouvertes, aussi appelées « verbatims » ou « verbatims clients ». Dans les méthodes qualitatives en sciences sociales, la codification est une tâche habituellement manuelle qui vise à organiser les verbatims en leur attribuant une « classe » ou un « code » c'est-à-dire un thème abordé dans le message ; les verbatims peuvent être multiclassés, c'est-à-dire qu'on peut leur attribuer plusieurs codes. L'ensemble des codes constitue un *plan de code* ou *plan de classement* (voir **Figure 10**).

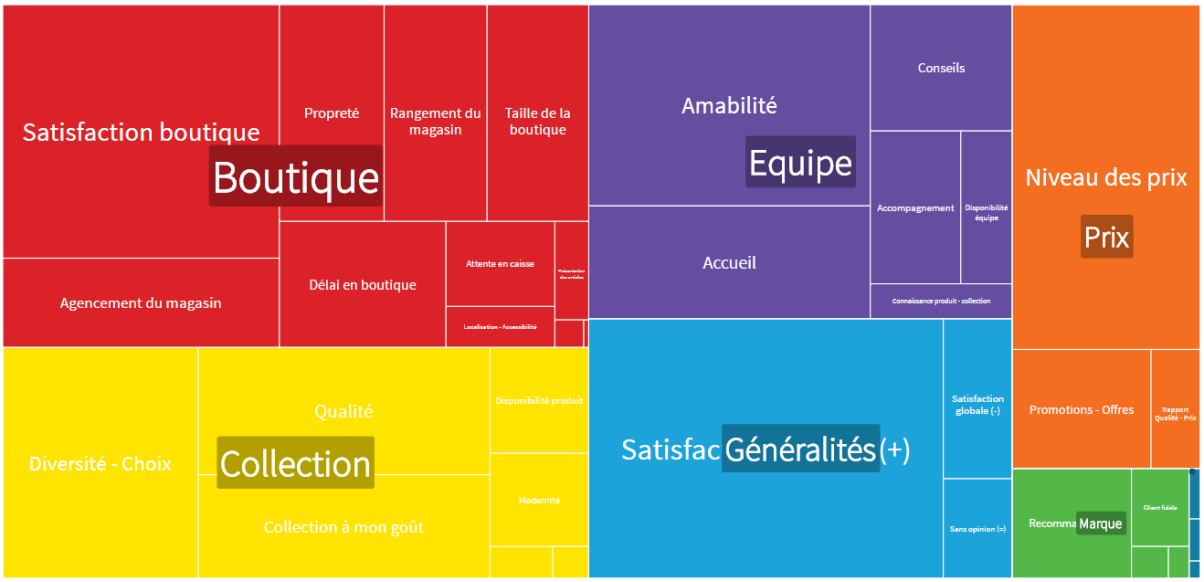

Figure 10. *Exemple de visualisation de type « treemap » d'un plan de classement. Les codes sont libellés et hiérarchisés dans une organisation à deux niveaux. Sur la visualisation, la surface couverte par chaque zone représente le nombre de verbatims correspondants.*

Par exemple le verbatim « Une bonne équipe plein de bon conseil » [sic] pourra être catégorisé avec les codes *Qualité du personnel* et *Qualité des conseils*. En répétant cette catégorisation sur chaque verbatim, la codification permet d'étudier l'ensemble des réponses d'une enquête ou d'un dispositif, en déterminant la fréquence des codes. Au terme de cette opération, le code devient une information quantitative qui peut être croisée avec d'autres variables, dont on peut mesurer l'évolution et qu'on peut visualiser dans un tableau de bord.

3. De la démarche représentative au projet « *big data* »

Les responsables de la relation client qui viennent, par leur parcours ou leur formation, des études de marché, sont particulièrement sensibles aux modalités de sondage des répondants, et notamment à la formulation de ces questions. Ils considèrent que deux questions formulées différemment mesurent des réalités distinctes, en accord avec la tradition méthodologique des sondages d'opinion et des enquêtes quantitatives en sciences sociales. Ils sont également sensibles à la représentativité de l'enquête qui doit, par échantillonnage, proposer un reflet fidèle de l'ensemble des clients de l'entreprise ou de son marché (constitué des clients actuels mais aussi potentiels, les prospects, et des non-clients). Nous avons retracé au chapitre IV la façon dont cette représentativité est considérée dans le monde académique et en entreprise : pour les responsables de la relation client également, elle détermine la valeur épistémique attribuée aux données. Les données d'enquêtes de satisfaction analysées par Proxem héritent donc au premier abord d'une épistémologie de la donnée qui provient de l'univers des enquêtes d'opinion. Les biais de sélection, d'échantillonnage, ceux introduits par les non-réponses, sont connus de ces responsables de la relation client ; ils ne sont pas attribués à Proxem et constituent une limite connue et admise de ses clients, qui s'opère en amont de sa contribution.

Par ailleurs, l'exploitation de ces enquêtes est une utilisation secondaire de données marquée par une rupture épistémique entre les conditions de constitution des données et les conditions d'analyse. L'analyste Proxem, comme le chargé d'études d'un institut de sondage, est toujours déjà privé de la connaissance des conditions initiales, contrairement au sondeur qui collecte l'information. Idéalement, les données sont documentées (le plus souvent de façon informelle et oralement par le client de Proxem) pour reconstituer le contexte de l'enquête d'un point de vue spatio-temporel, méthodologique et épistémique. Néanmoins cette transmission orale peut être incomplète et soumise, comme toute communication, à des incompréhensions. La reconstitution du contexte épistémique de l'enquête facilite, en fin de processus, l'interprétation des données calculées par Proxem, mais n'est pas toujours possible.

De manière générale, la population complète des clients est difficilement accessible : les outils de gestion de la relation client (*Customer Relationship Management*, ou CRM) s'appuient sur des bases de données constituées de manière souvent erratique, par des intervenants successifs, sur la base d'agrégations qui génèrent souvent des pertes d'information, des doublons et des informations manquantes : par exemple, pour un client donné, on aura l'historique de ses achats, et pas pour un autre, dont on connaîtra en revanche l'âge, ou le magasin préféré, etc. Les clients disposant d'une carte de fidélité sont souvent surreprésentés, car ils doivent généralement renseigner des informations pour obtenir la carte, ce qui n'est pas le cas de la population générale des clients ; le fait qu'ils aient choisi de prendre une carte de fidélité en fait évidemment une population spécifique, non représentative de la population générale, plus fortement attachée à l'enseigne[37]. Les enquêtes à chaud ont pour population de référence les clients qui ont réalisé un achat sur une période de temps spécifique, et non l'ensemble des clients de l'entreprise. Certaines enquêtes, à chaud comme à froid, ne sont mises en place que pour les clients du site de commerce électronique, et non en magasin, et ne sont donc représentatives, au mieux, que de ce sous-ensemble des clients. Les enquêtes à chaud sont critiquées pour les réactions plus extrêmes, plus émotionnelles, qu'elles peuvent susciter ; les enquêtes à froid sont quant à elles réputées imprécises, plus difficiles à interpréter. De manière générale, toutes ces spécificités liées au mode d'enquête, et reconstituables par documentation et recontextualisation, sont en pratiques écrasées par l'agrégation de ces différentes enquêtes dans les bases de données CRM.

Cette agrégation a des conséquences importantes sur la valeur épistémique des données : si la représentativité de chaque type d'enquête est difficile à reconstituer individuellement, la représentativité de l'agrégation de toutes les enquêtes constituées selon des normes hétérogènes est

[37] Contrairement à ce que l'on pourrait penser, la satisfaction n'est pas toujours plus élevée chez les clients fidèles, car, corrélativement à leur attachement à l'enseigne, ils ont également souvent des attentes plus élevées, et sont plus sensibles aux déceptions que peut leur causer l'enseigne.

tout simplement inaccessible. Il n'existe pas de procédure qui permet de l'estimer, de la reconstituer, pour un agrégat donné, si ce n'est en redécoupant la base de données en unités homogènes.

Cette agrégation est donc un point crucial de rupture par rapport à la continuité épistémique du traitement de données textuelles. Elle est mise en avant par Proxem comme justification de son positionnement d'acteur des *big data* traitant des données variées. C'est un point de rupture du point de vue des normes des sondages d'opinion, mais aussi plus spécifiquement par rapport à la codification traditionnelle, qui valorise et s'appuie sur une homogénéité au moins présumée du corpus de données textuelles. En effet, un corpus textuel doit traditionnellement être « régulé sur le plan linguistique » et présenter « une forte stabilisation thématique » (Cailliau et Poudat 2008) pour rendre possible une approche statistique du texte (telle qu'en lexicométrie).

Dans l'approche agrégée suivie par Proxem, la stabilisation thématique ne vient pas de l'unicité de la source mais de la ressemblance entre les contextes pragmatiques des données textuelles issus de sources diverses : l'agrégation se justifie par le fait que le contexte d'énonciation du verbatim est celui d'un client qui cherche à faire part d'une remarque ou d'un problème à une entreprise. Cette intention de communication est ce qui permet de rendre les verbatims commensurables entre eux, quelle que soit leur origine. L'homogénéité ne provient donc pas du mode de constitution des données elles-mêmes, mais de leur agrégation comme geste épistémologiquement constitutif de l'ensemble, qui permet malgré tout à l'agrégat de faire corpus. Cet axe de commensuration sera par la suite l'axe d'interprétation des données, qui seront analysées du point de vue de cette relation communicationnelle entre le client et l'entreprise. La justification de cette commensuration est empirique : que le client s'exprime par mail, par écrit ou par téléphone, Proxem a constaté qu'en pratique les thématiques abordées sont suffisamment homogènes pour pouvoir appliquer le même plan de code sur des sources différentes agrégées en corpus. Cette commensuration est également possible d'un projet à l'autre : les thématiques identifiées sont suffisamment stables d'une entreprise à l'autre et d'un secteur d'activité à l'autre pour permettre un recouvrement au moins partiel. Il est alors possible de comparer les résultats obtenus pour un client de Proxem à d'autres résultats antérieurs pour constituer une norme empirique à l'échelle d'un secteur d'activité, et non plus à l'échelle d'un corpus unique.

4. La préparation des données

Le geste épistémologiquement constitutif d'agrégation et de commensuration des données se double d'un ensemble de gestes techniques, par lesquels les données textuelles sont effectivement rendues manipulables par les salariés de Proxem. Plus spécifiquement, le traitement des données est réalisé par une équipe d'infolinguistes, c'est-à-dire de personnes qui possèdent des compétences techniques en informatique et en linguistique. Ces personnes sont capables *a minima* d'utiliser des logiciels d'analyse linguistique mais aussi de développer ou faire évoluer de tels logiciels. Au sein de

l'entreprise, elles ont également une mission de chefs de projet, c'est-à-dire qu'elles sont en charge de comprendre les attentes du client, répondre à ses demandes, organiser et respecter un planning de travail, et enfin veiller à la qualité des analyses.

Parmi les opérations préalables au traitement linguistique des données, la première étape de leur travail est l'import des données. Cette étape est loin d'être purement technique car elle dépend d'un certain nombre de choix ponctuels (effectués au moment d'un projet spécifique) et de choix systématiques (effectués au moment de la conception de l'outil, et par conséquent communs à tous les projets). Il est en effet nécessaire de faire entrer les données fournies par le client dans le modèle pratiqué par Proxem ; cette étape correspond à la phase de nettoyage considérée comme cruciale dans les sciences des données.

Les données sont généralement reçues sous la forme d'un fichier Excel où chaque ligne correspond à un individu et chaque colonne à une ou plusieurs questions. Il arrive souvent que les questions soient renommées ou représentées par un code ; l'infolinguiste en charge du projet doit alors « décoder » le fichier pour reconstituer les questions et les réponses et y remettre du sens. Parfois, le codage n'est pas cohérent : par exemple, sur une partie du tableau, le genre est codé par « M » ou « F », et sur autre partie en « 1 » ou « 2 » ; il n'est pas rare que d'autres chiffres ou lettres (comme « G » ou « H », qu'on pourrait réinterpréter, mais aussi « B, « K », « 00 », etc.), sans sémantique attribuée, figure dans une même colonne « genre ». C'est alors à l'infolinguiste de décider si elle[38] peut recoder ces valeurs aberrantes, ou doit les convertir en valeurs manquantes. Comme dans tout traitement de données, la gestion des valeurs manquantes est problématique : ainsi, dans notre exemple, on pourra calculer le ratio homme/femme uniquement sur les individus dont on a pu déterminer le genre ; si l'on veut comparer de quoi parlent hommes et femmes dans l'enquête, on travaillera alors sur un sous-échantillon des données qui ne représente pas 100% des avis exprimés. Ces valeurs manquantes sont un problème qui apparaît dès l'acquisition des données, ne peut être résolu par l'infolinguiste, et nuit à l'exactitude de l'analyse ultérieure.

Lorsque les questionnaires sont longs, l'infolinguiste peut également être amenée à sélectionner un sous-ensemble des questions fermées jugées plus intéressantes pour le projet : cette sélection fait l'objet d'une négociation opérationnelle et commerciale, le nombre maximum de colonnes prises en compte pouvant être spécifié dans le contrat qui lie le client à Proxem. Ces questions fermées sont ensuite converties en métadonnées des verbatims, c'est-à-dire des informations qui caractérisent l'auteur du verbatim. La façon dont ces métadonnées sont ensuite organisées fait également l'objet de discussion entre l'infolinguiste et le client, ou se fait en interne selon la disponibilité des différents interlocuteurs. Par exemple, il n'est pas rare qu'il y ait une ou plusieurs colonnes de localisation géographique avec des granularités variables (code IRIS INSEE, ville, code postal, département, région,

[38] A ce jour, l'équipe d'infolinguistes de Proxem est entièrement féminine.

etc.) dont certaines propres à l'enseigne (directions régionales, magasin, zone géographique théoriquement couverte par le magasin, etc.) qui doivent être réconciliées, souvent au prix d'une perte d'information : par exemple, si l'on dispose pour certains individus de leur code postal, et pour d'autre de leur département, il faudra tout ramener au département pour avoir des informations commensurables. Les catégorisations apparaissant de manière hiérarchique dans l'outil de Proxem, il est aussi possible de préserver ces différents niveaux de granularité : ce sera alors à l'utilisateur du logiciel de veiller à ne pas comparer des éléments suivant des granularités différentes. Ces problèmes d'hétérogénéité des données peuvent se poser au sein d'un même fichier, notamment lorsqu'il a été constitué manuellement par des personnes différentes, et plus encore lorsque le projet repose sur l'agrégation de plusieurs sources de données.

Dans le traitement d'enquête traditionnel également, il y a toujours déjà une rupture épistémique due à la division du travail. La personne qui interroge les sondés et retranscrit leurs réponses n'est pas celle qui conçoit le questionnaire, ni celle qui en fera la codification. Certains instituts de sondage externalisent l'administration effective du questionnaire : ce sont non seulement des personnes différentes, mais appartenant à des entreprises différentes, qui font la collecte et le traitement des données. Il y a alors un aller-retour entre l'institut, qui conçoit le questionnaire, puis en analyse les résultats, et le prestataire qui réalise le terrain. Comme le souligne Rémy Caveng (2012) pour le champ académique :

> [L]es utilisateurs d'enquêtes quantitatives ne se posent guère la question des conditions réelles de recueil des données qu'ils traitent. Cette « indifférence » au terrain qui contraste avec les injonctions à la réflexivité critique considérée comme dimension constitutive de l'expérience ethnographique et de sa validité scientifique (Beaud, Weber, 1998 ; Olivier de Sardan, 1995 ; Schwartz, 1993) s'explique assez aisément. L'exploitation d'une enquête quantitative consiste en la manipulation de fichiers d'où les conditions de recueil de l'information sont absentes. Le type de traitements ainsi que la division du travail propre à la production de ces enquêtes tendent à occulter l'activité qui, entre la conception des questionnaires et leur saisie, a été déployée pour obtenir un résultat qui se présente comme un ensemble inerte de lignes et de colonnes.

Cette division du travail aboutit à une considération quasi-nominaliste de l'enquête : l'opinion n'est pas ce qui a été recueilli sur le terrain, mais ce qui est matérialisé dans les verbatims. Le terrain réel est inconnu, rendu inaccessible par le dispositif d'enquête et la division du travail. Pour l'analyste ou le codeur, l'observable qu'il devra manipuler n'est pas l'opinion elle-même, mais l'image qu'en propose le verbatim. Néanmoins, cet état de fait n'aboutit pas à une dévalorisation des données où elles seraient reléguées au rang de construction arbitraire inexploitable. Au contraire, la positivité du verbatim (plutôt que du terrain) donne lieu à une épistémologie positiviste, renforcée par l'usage de certains logiciels de traitement de données qualitatives (conçus pour la codification de verbatims), qui ne donnent pas à voir un « ensemble inerte de lignes et de colonnes », mais des éléments déjà pré-organisés, pré-sémantisés.

Il faut noter également que dans les enquêtes qualitatives, où l'analyste est généralement aussi le sondeur et le transcripteur, cette rupture entre le terrain et l'analyse n'existe pas, ce qui permet une réflexivité critique par rapport au terrain et explique au moins en partie une épistémologie plus classique des sciences humaines telle que l'ethnométhodologie ou l'interactionnisme (Fielding et Lee 1997). Ces approches fondées sur le fait que l'objet observé est bien la personne et non le verbatim écrit sont difficilement soutenables dans le cas des enquêtes quantitatives, notamment en raison de cette division du travail entre sondeur et analyste, et du cadre méthodologique spécifique des questionnaires de sondage. Ainsi, bien que les traitements proposés par Proxem portent sur des données traditionnellement vues comme qualitatives (du texte), l'épistémologie qui les régit est résolument celle des enquêtes quantitatives avec leur positivisme de la donnée, du fait d'une division du travail similaire. Lorsque l'infolinguiste reçoit et importe les verbatims, elle est toujours déjà privée d'un accès direct au réel, qui ne peut être que partiellement reconstitué par recontextualisation. Néanmoins, la seule attitude soutenable pour l'activité épistémique des infolinguistes est d'envisager les verbatims dans leur positivité et de s'appuyer sur cette positivité pour constituer des connaissances. Si, au terme de l'analyse linguistique, il est possible de remettre en cause la validité des connaissances produites au nom de la qualité insuffisante des données, cette qualité est pour l'infolinguiste un donné qui ne peut pas être modifié ni négocié du point de vue de sa valeur épistémique, et constitue le point de départ de l'analyse.

5. La conception du plan de code

Après que les données ont été formatées, le travail principal de l'infolinguiste consiste à étudier les verbatims pour en faire ressortir les principales thématiques et structurer ces thématiques dans un plan de code. Cette étape est particulièrement délicate : selon la façon dont les classes sont délimitées, elles rassembleront plus ou moins de verbatims et seront donc plus ou moins importantes dans les classements. Les infolinguistes de Proxem considèrent qu'il n'existe pas un plan de code idéal dont on chercherait à se rapprocher, et qu'un même ensemble de verbatims peut être couvert par des plans de codes distincts et qui ne se ressemblent pas forcément. L'épistémologie de la construction du plan de code chez Proxem est ainsi similaire à celle des enquêtes quantitatives : il s'agit d'un va-et-vient entre un principe d'empiricité visant à faire émerger les thématiques des données (approche ascendante), et un principe d'ordre rationnel visant à dégager une structure cohérente de thématiques (approche descendante). La construction d'un plan de code est un exercice délicat de négociation entre ces principes qui doivent être appliqués avec sagacité.

D'une part, le plan de code est une construction sociale et théorique qui n'a pas d'optimum absolu et naturel : en contradiction avec l'ambition de Socrate dans le *Phèdre* de découper une idée « selon ses articulations naturelles », il n'existe aucun système de catégorisation qui découpe parfaitement le domaine concerné par les verbatims. L'approche descendante est ainsi considérée comme un risque

particulièrement grave pour l'analyse d'enquêtes. L'utopie de la catégorisation a largement été analysée en sociologie de la quantification en ce qui concerne notamment la classification des maladies (Bowker et Star 1999), la taxonomie phylogénétique (Desrosières 2010) ou encore les nomenclatures socioprofessionnelles (Amossé 2013). Ce sont autant d'exemples de typologies historiquement constituées et plus ou moins stabilisées, bien que régulièrement remises en cause, qui ne peuvent pas être abordées avec un regard naturaliste ou positiviste, mais constituent néanmoins des conventions utilisables et utilisées.

D'autre part, l'idée de faire émerger « naturellement » des catégories à partir de la matière empirique est elle aussi une utopie : toutes les techniques de classification non supervisée donnent des résultats imparfaits qui font ressortir un mélange de catégories pertinentes et de catégories incompréhensibles, c'est-à-dire des résultats souvent hétérogènes et difficiles à interpréter. Par exemple, les infolinguistes de Proxem s'appuient sur des outils d'extraction terminologique qui font ressortir les mots et expressions les plus fréquentes à l'intérieur d'un ensemble de verbatims ; or, sur des corpus d'analyse d'emails, les expressions qui dominent quantitativement sont généralement les formules de politesse, qui ne renvoient à aucune thématique d'un plan de code. Les résultats de cet outil ne peuvent pas être exploités tels quels : en revanche, ils sont une aide précieuse pour l'infolinguiste qui s'emploie à faire émerger les principales thématiques d'un corpus. Comme le disent également Aurélien Bénel, Christophe Lejeune et Chao Zhou (2010), les informations qui se dégagent

> ne constituent en rien une « analyse automatique » mais plutôt une invitation à la lecture attentive. Leur raison d'être est heuristique : une saillance quantitative peut suggérer à l'analyste l'ouverture d'une investigation.

La tension entre les deux principes de constitution du plan de code, sur laquelle nous allons revenir, s'accompagne de conditions pratiques communes à l'analyse traditionnelle d'enquêtes quantitatives avec questions ouvertes et le travail proposé par Proxem. L'analyste gagne naturellement à être familier du domaine analysé : chez Proxem, les infolinguistes développent avec le temps des spécialisations dans les expertises proposées par l'entreprise, dont l'expérience client. Comme on l'a vu, il y a une stabilité des thématiques abordées dans ce contexte qui permet de se constituer un horizon d'attente raisonnablement réaliste en première approche. Il existe par ailleurs quelques règles plus ou moins tacites de ce qui constitue un bon code en analyse de verbatims : il est facile d'y rattacher des verbatims, il décrit un aspect concret de l'opinion des sondés, il est peu ambigu, il se distingue bien des autres codes. Des conventions relatives au plan de code existent également : les codes doivent être de taille comparable (coder un nombre similaire de verbatims) et ne laisser que peu de verbatims « autres » n'appartenant à aucun code. Xavier Marc (2001) souligne que l'on

> veille notamment à conserver un poste "autre" inférieur à 5% ; on veille également à éviter les codes qui ne correspondraient qu'à 2 ou 3% des réponses.

Par ailleurs, dans le cas d'un recueil en face-à-face, certains sondeurs ont tendance à pré-codifier les réponses en les reformulant d'une façon qui va vers l'uniformisation de l'expression et met en évidence les classes du plan de code que le sondeur a déjà en tête. Dans cette configuration, la division du travail est moins nette, mais l'analyste ne sait pas forcément à quel point le sondeur a déjà reformulé le propos exprimé. Une disparité souvent remarquée à ce stade est que certains sondeurs retranscrivent la réponse en discours indirect libre (« il est déçu de la qualité des matériaux utilisés ») plutôt qu'en discours direct (« je suis déçu de la qualité des matériaux »). Si plusieurs sondeurs ont fait la collecte, et que certains ont adopté cette pratique, les verbatims seront plus hétérogènes, plus difficiles à analyser de manière semi-automatique.

Pour les modes d'administration du questionnaire où le sondé s'exprime directement sans retranscription, notamment lorsqu'il le remplit lui-même par écrit, cette reformulation n'a évidemment pas lieu, ce qui va *a priori* vers une plus grande authenticité du propos recueilli : dans ce contexte, une donnée « brute » est une réponse non reformulée par un sondeur. Néanmoins, ce n'est pas parce que le sondeur n'a pas précodifié les réponses que l'analyse ne risque pas de reposer, avant même d'avoir commencé, sur un plan de code trop rigide et trop éloigné de la réalité des réponses. L'existence d'un matériau déjà écrit ne protège pas des approches trop descendantes de la codification (où l'analyse prédéfinit des codes et va ensuite les rechercher dans le texte) qui peuvent venir de l'horizon d'attente de l'analyste, de l'usage prescrit par la façon dont le logiciel est conçu (Bénel, Lejeune et Zhou 2010) ou de contraintes extérieures prescrites par un client par exemple. Si les *a priori* de l'analyste peuvent être révisés, il est plus difficile de modifier le logiciel qui génère des pistes de catégories, ou de faire changer d'avis le client qui considère une catégorie comme cruciale.

La constitution du plan de code est donc un art, associé à un savoir-faire, qui mobilise un ensemble de compétences linguistiques mais aussi extralinguistiques. Dans les enquêtes traditionnelles, l'analyste est maître à la fois du plan de code et de la façon dont les verbatims y sont répartis. Les frictions, les hésitations, les décisions éventuellement discutables qui sont faites lors de la codification manuelle ne sont presque jamais mises en évidence, ni dans le monde universitaire, ni dans les instituts de sondage : le plan de code est un donné de l'analyse quantitative. Il y a là un facteur d'opacité où la subjectivité de l'analyste est censée être compensée par sa maîtrise des principes de la codification ; dans les faits, cette maîtrise est supposée mais non testée. L'utilisation des outils d'analyse sémantique de Proxem oblige à lever une partie de cette opacité. D'une part, comme on l'a vu, le plan de code est le fruit d'une négociation entre le client et l'infolinguiste et n'émerge qu'au terme de discussions qui peuvent être longues et être réactivées plusieurs mois avec le début effectif du projet : cela veut dire que le plan de code n'est pas livré au client comme un état de fait, mais construit avec lui.

D'autre part, comme nous allons maintenant le voir, la mécanisation de la codification oblige à exprimer explicitement la logique de la codification sous la forme de règles formelles ou semi-

formelles qui seront appliqués automatiquement par le logiciel. En utilisant le logiciel, le client peut voir dans chaque verbatim quelle expression active telle ou telle classe du plan de code, signaler les erreurs ou contester certaines activations. Comparée à la codification manuelle faite dans les instituts de sondage, la démarche adoptée par Proxem apparaît de ce point de vue comme plus transparente et plus collaborative.

6. La constitution des ressources linguistiques

Le classement effectif des verbatims dans le plan de code ne se fait pas unitairement, mais à travers la composition de règles linguistiques fondées sur la présence ou l'absence de certains mots, leur ordre, leur nature, etc. Il y a donc un travail de systématisation qui consiste à identifier les éléments lexicaux justifiant qu'un verbatim soit rattaché à un code. Dans le logiciel, les segments de texte qui activent les règles linguistiques sont soulignés : on parle d'*annotations* sur le texte. Ces annotations portent sur des *entités nommées*, un terme conventionnellement utilisé en linguistique computationnelle, mais qui porte habituellement sur des éléments sémantiques spécifiques, à savoir certains nombres propres : les personnes, les lieux et les organisations. L'annotation renvoie aux observables lexicaux du texte, là où l'entité nommée constitue son libellé. De ce fait, le segment de texte souligné peut être différent de son libellé. Chez Proxem, les annotations peuvent être des noms communs, des verbes, des adjectifs, mais aussi des groupes de mots de nature variée. Grâce aux annotations, l'utilisateur qui parcourt les verbatims peut consulter non seulement les codes activés, mais aussi les entités nommées qui ont activé ces codes.

Cependant, les règles de codification appliquées par le logiciel ne sont pas entièrement transparentes ; leur écriture et leur compréhension relève d'une technicité qui n'est que rarement transférée au client. De plus, la taille des corpus peut décourager le client d'une vérification fine, voire unitaire, de l'analyse, et oblige à une distanciation par rapport aux documents qui peut être rapprochée de la distinction entre lecture proche et lecture distante en analyse littéraire (Moretti 2000).

```
1  élevé→0.9→:trop cher→+prix→+trop
2  attractif→0.9→:pas cher→+prix
3  augmenter→0.9→:plus cher→+prix
4  bémol→0.9→:trop cher→+prix
5  casser→0.9→:moins cher→+prix
6  cher→0.9→:trop cher→+devenir→+vraiment→-pas
7  comparateur→0.9→:comparateur de prix→+prix
8  comparateur de prix
9  comparatif→0.9→:comparateur de prix→+prix
10 comparer→0.9→:comparateur de prix→+prix
11 excessif→0.9→:trop cher→+prix
12 honéreux→:trop cher
13 hors de prix→:trop cher
14 kilo→0.9→:prix au kilo→+prix→+euro
15 même prix
16 meilleur prix→:moins cher
17 meilleur rapport qualité prix→:moins cher
18 mm prix→:même prix
19 moins cher
20 pas cher
```

Figure 11. *Les 20 premières entrées de la gazette* Prix *utilisée pour les projets dont le client appartient à la grande distribution. On retrouve le score déterminant si l'observable lexical doit être détecté, des redirections, des activateurs et des inhibiteurs. Ainsi, sur la première ligne, l'adjectif « élevé » sera annoté pour la notion de prix s'il est précédé ou suivi du mot « prix » ; cette annotation renverra vers une entité nommée « trop cher ».*

Jusqu'en 2016, où un nouveau logiciel d'écriture de règles linguistiques est progressivement mis en place[39], les règles de codification s'écrivent sous la forme d'entrées dans des dictionnaires spécialisées, les *gazettes*[40], qui prennent la forme de fichiers texte (**Figure 11**). Une entrée est composée de plusieurs éléments séparés par des tabulations. L'élément fondamental est une expression composée d'un ou plusieurs mots, généralement ramenée à sa forme de base, aussi appelé lemme (infinitif pour les verbes, masculin singulier pour les noms et adjectifs). Lorsqu'un élément est écrit sous la forme d'un lemme, le moteur d'analyse va prendre en compte toutes les formes fléchies de cet élément. Ainsi, si l'infolinguiste crée l'entrée « acheter » dans une gazette, les mots « acheté », « achète », « achèterai », seront détectés. À l'inverse, si elle crée l'entrée « achèterai », seule cette forme sera analysée. D'autres éléments peuvent être ajoutés à l'entrée principale :

- un score entre 0 et 1, qui décrit la probabilité *a priori* que l'expression doive être reconnue ;
- des mots "activateurs" ou "inhibiteurs" (respectivement précédés d'un symbole "+" ou "-") qui influent sur le score pour l'augmenter ou le diminuer ;
- depuis 2016, des indications « left » ou « right » qui permettent de spécifier si l'activateur ou l'inhibiteur doit obligatoirement se trouver à gauche ou à droite de l'entrée ;

[39] Plus précisément, deux évolutions se sont produites par rapport à cette configuration de départ. Courant 2015, un logiciel a été développé en interne, qui permet de centraliser la base de connaissances linguistiques et de la modifier *via* un logiciel interne, installé sur le poste des infolinguistes, plutôt que de gérer cette base de connaissances avec un système de gestion de version (versioning). En mars 2017 est lancé un logiciel qui permet non plus uniquement aux infolinguistes de Proxem, mais également à ses clients, de créer et manipuler des règles linguistiques, revisitées avec une nouvelle syntaxe.
[40] De l'anglais 'gazetteer', qui peut être traduit littéralement par « répertoire », « nomenclature » ou « index ».

– une autre expression vers laquelle l'entrée doit être redirigée pour former un regroupement.

Parmi tous ces éléments, seule la première expression est obligatoire ; les autres éléments servent à contourner des difficultés d'ordre linguistique :

– l'ambiguïté et la polysémie des mots ou expressions ;
– la synonymie entre termes ;
– les relations d'hyponymie et d'hyperonymie entre concepts (par exemple, dire qu'une orange est un fruit fait de « orange » l'hyponyme de « fruit », et réciproquement, de « fruit » l'hyperonyme de « orange »).

Par ailleurs, l'infolinguiste qui prépare ces entrées doit anticiper ou désactiver au cas par cas les traitements automatiques de normalisation qui y sont appliqués : suppressions des diacritiques (accents, trémas, cédilles, etc.), racinisation, minusculisation. Il est ainsi possible, par exemple, d'obliger le moteur d'analyse à ne détecter un terme que s'il est écrit en majuscule, ou accentué. Sans spécification, la phase de normalisation augmente la performance et la généralité de l'outil au prix d'une perte d'information qui est significative du point de vue de la linguistique mais considérée comme négligeable d'un point de vue sémantique et pragmatique : l'enjeu du moins est, pour effectuer les calculs dans une durée raisonnable, de « réaliser une réduction de données avec une "perte minimale de sens" » (D'Aubigny 2001). Dans l'ensemble, les infolinguistes évaluent au cas par cas la solution optimale.

Notons par ailleurs que l'écriture de ces entrées ne se fait pas intégralement à partir de zéro pour chaque projet : un certain nombre de gazettes ou d'entrées peuvent être réutilisées telles quelles ou après ajustement. L'infolinguiste peut puiser dans une base de connaissances linguistiques constituée de manière collaborative et cumulative au fil des projets. Dans le domaine de l'expérience client, le recours à cette base de connaissances peut être la principale source d'entrées. Lors de projets similaires du point de vue de leur contexte linguistique et pragmatique (secteur d'activité du client, longueur et mode de constitution de données), les entrées d'un projet peuvent être réutilisées telles quelles ou presque sur l'autre projet : c'est notamment un moyen efficace en termes de ressources de proposer une démonstration du logiciel à un client potentiel sur ses propres données, en précisant que l'analyse proposée présente une marge d'amélioration. Les projets relatifs à l'expérience client disposent ainsi non seulement d'entrées, mais d'un plan de code intermédiaire constitué de points positifs (dits *Félicitations*) et négatifs (dits *Problèmes*) récurrents d'un projet à l'autre, qui sont ensuite réorganisés en un plan de code co-construit par le client et l'infolinguiste à l'occasion du projet.

Les projets ne sont d'ailleurs pas le seul moteur de constitution de gazettes : les infolinguistes s'appuient également, lorsque c'est possible, de bases de connaissances préconstituées qui pourraient être importées dans le formalisme des gazettes, *a minima* sous la forme d'une liste à plat. Certaines gazettes, comme les noms de marques, les toponymes, les noms de couleurs, etc., sont stables d'un

projet à l'autre et les expressions analysées restent quasiment identiques. Par exemple, les noms des départements français sont globalement stables. Cependant, d'autres formulations, d'autres périphrases, peuvent exister, et certains peuvent avoir des homonymes. Ainsi, le Cantal est un département mais c'est aussi un fromage : dans un projet dont le contexte pragmatique est celui de la grande distribution, il faudra retravailler l'entrée *Cantal* dans la gazette des départements en ajoutant des activateurs ou des inhibiteurs car la mention du fromage est plus probable que celle du département.

Une manière d'alimenter ces gazettes majoritairement stables est de construire des ponts avec les ontologies du web sémantique. Proxem a constitué sa propre ontologie à usage interne à partir de Wikipédia (indépendamment du projet DBPedia) et est aussi contributeur de la base de données lexicales WordNet en français. À l'inverse, d'autres techniques peuvent être mobilisées pour générer automatiquement « à la volée » des gazettes ou des éléments de gazettes spécifiques au projet. L'extraction terminologique mentionnée plus haut est l'un des moyens de faire cette opération. Proxem intègre également depuis 2016 les *word embeddings* de Mikolov et al. (2013), une technique qui permet de générer une base d'expressions multi-mots à partir d'un corpus et représentées formellement non pas comme des « sacs de mots » (où les mots d'une expression sont stockés « à plat », sans hiérarchie ni structuration) mais comme des vecteurs permettant des opérations algébriques d'addition et de soustraction sur des expressions. Bien que fondée sur des avancées technologiques récentes à base de réseaux de neurones artificiels, cette approche s'inscrit dans la logique de la linguistique distributionnelle à la Harris et Bloomfield (Sahlgren 2008) où le sens d'un mot dépend de son contexte d'usage, et notamment des relations de cooccurrences qu'il entretient avec les autres mots, en particulier les mots proches. Le distributionnalisme renvoie à une conception du langage plus proche de Wittgenstein que du Cercle de Vienne. Après la textométrie (Lebart et Salem 1994) et l'analyse sémantique latente (Landauer et Dumais 1997), l'approche distributionnelle continue ainsi à servir de cadre théorique (souvent ignoré des informaticiens et statisticiens) aux approches statistiques et computationnelles du langage.

Pour ce qui est de l'acquisition et de la constitution de ressources linguistiques, l'épistémologie qui régit le travail des infolinguistes est donc plutôt un *anything goes* à la Feuerbach où tout ce qui peut contribuer à améliorer la finesse et la couverture des annotations sémantiques est intégré, quitte à être remanié ultérieurement pour s'homogénéiser avec le reste des ressources. La question du statut épistémique de ces ressources ne se pose donc pas à ce stade, contrairement aux débats qui peuvent avoir lieu parmi les concepteurs d'ontologies. Il n'est pas question de constituer une sémantique formelle où le sens se modélise comme un calcul, ni une représentation universelle des concepts avec une visée réaliste, mais d'identifier un maximum d'éléments lexicaux structurés en taxonomies et pouvant se rapporter à des concepts susceptibles d'intéresser le client. Le sens des observables produits par l'analyse sémantique relève plutôt, comme en textométrie, d'une sémantique

interprétative à la Rastier (Pincemin 2012), mais désormais collaborative ou sociale puisqu'y interviennent à la fois l'infolinguiste et le client. De plus, la démarche relève davantage de l'ingénierie, de la construction effective de ressources linguistiques, que de leur validation scientifique.

7. L'alimentation du plan de code

Dans les deux parties précédentes, nous avons décrit successivement deux étapes présentées comme relativement distinctes : la constitution du plan de code et la création de ressources linguistiques. En pratique, il y a un va-et-vient permanent entre ces deux tâches. En effet, l'identification d'entités nommées motive la création ou l'évolution d'un code, tandis qu'un travail plus conceptuel de structuration des thématiques fait émerger des codes auxquels il faut trouver des correspondants lexicaux. Par ailleurs, il y a une étape intermédiaire entre la détection d'entités nommées et l'alimentation du plan de code. Cette étape permet de combiner, soustraire ou fusionner des entités nommées avant de les transformer en code. Par exemple, une enseigne de grande distribution peut détenir plusieurs centaines de magasins, souvent identifiés par ses clients *via* leur emplacement géographique. Pour détecter ces magasins, les infolinguistes génèrent des codes correspondant à la cooccurrence dans un segment de phrase de l'entité nommée « magasin » et des toponymes (noms de lieux). Ainsi un verbatim comme « je me suis rendu dans l'hyper de Pontoise pour acheter du cantal » activera en principe des codes comme *visite en magasin*, *magasin de Pontoise* et *fromage*. On notera en passant que certains codes ne sont pas systématiquement des thématiques comme on les voit classiquement dans un plan de code. En réalité, plusieurs plans de code avec des éléments de nature assez différente peuvent coexister au sein d'un même projet.

Du fait de ce travail itératif, la volonté d'identifier des éléments lexicaux signifiants et linguistiquement valides n'est pas la seule motivation de l'infolinguiste. Ce n'est pas un travail purement linguistique visant à rendre compte de la réalité lexicale et sémantique du corpus : ce travail est instrumentalisé comme il l'est en lexicométrie, discipline d'ordre linguistique au service des sciences sociales (Guérin-Pace 1997). La recherche d'éléments lexicaux significatifs est infléchie par d'autres motivations :

- la recherche d'un consensus avec le client qui a à l'esprit certaines thématiques qu'il souhaite spécifiquement identifier ;
- la réutilisation des ressources linguistiques déjà existantes autant que possible et raisonnable ;
- la prise en compte d'un maximum d'éléments du corpus dans l'optique d'annoter au moins une fois chaque verbatim.

L'alimentation du plan de code est donc une négociation interne et externe : interne, car l'infolinguiste arbitre et fait des choix modelés par des injonctions différentes. Ainsi le fait de réexploiter des ressources existantes peut être motivé par des contraintes de délais, par l'insuffisance du budget alloué

au projet, ou encore par la motivation personnelle de l'infolinguiste. La négociation peut aussi être externe, avec le client, qui impose parfois un plan de code en amont du projet. L'infolinguiste teste alors le plan de code au regard des thématiques effectivement abordées dans les verbatims, et s'efforce de convaincre le client de revoir, reformuler ou supprimer les codes qui n'ont aucun ou peu de pendant empirique dans les verbatims. La reformulation peut faire l'objet de conversations prolongées. En effet, les libellés des codes doivent être parlants pour tous les interlocuteurs, décrire leur contenu de façon explicite, tout en correspondant à un vocabulaire « métier » (un sociolecte ou un technolecte, pour le dire en termes linguistiques) propre au client. Cette négociation parfois longue fait partie des éléments qui peuvent conduire l'infolinguiste à avoir envers le plan de code une attitude nominaliste où la classe ne correspond qu'à ce qui a été délimité comme tel. Du point de vue du client, c'est au contraire le moyen de parvenir à une posture réaliste dans laquelle les classes vont pouvoir recevoir une signification. Si le sens linguistique des classes et de leur contenu est, à quelques écarts près, commun à l'infolinguiste et au client, c'est surtout ce dernier qui peut anticiper les enjeux pratiques associés aux codes, nommément les problématiques qu'ils illustrent pour l'entreprise.

Une autre négociation importante se joue au niveau des regroupements, lorsque le plan de code est une cocréation de l'infolinguiste et du client. C'est souvent ce travail de regroupement et redécoupage qui suscite le plus de discussions. Par exemple, les responsables de centres de traitement de la relation client ont à gérer, dans leur quotidien, plusieurs types de délais : délai d'attente téléphonique, délai de réponse aux emails, délai de traitement du problème du client, délai de remboursement d'un achat, etc. Du point de vue de l'infolinguiste, la configuration la plus facile à mettre en œuvre est de créer un code unique *Délais*. Du point de vue du client, ces différents types de délais gagnent à être distingués : la gestion des emails et des remboursements peut appartenir à deux services complètement distincts au sein de l'entreprise, qui ne seront pas du tout intéressés par les remontées qui concernent l'autre service. En revanche le traitement des demandes par email et par téléphone peut appartenir au même service, ce qui permet de regrouper les deux notions de délais correspondantes. D'un côté, le client peut chercher à coller à la réalité de l'entreprise, ou tout simplement désirer que la codification soit la plus fine possible. À l'inverse, l'infolinguiste peut montrer qu'un verbatim comme « délai d'attente insupportable » est trop imprécis pour déterminer s'il s'agit d'un contact par téléphone ou par mail. Si l'information existe, le client pourra alors proposer d'enrichir les données fournies avec la source du contact, qui permettra de désambiguïser la notion de délai à partir d'une métadonnée. Par ailleurs, le client peut décider que du point de vue de la politique de l'entreprise, il faut montrer que les délais sont un problème important. Il demandera alors à l'infolinguiste de regrouper au maximum les différentes notions de délai, d'approfondir cette thématique à la recherche d'éléments lexicaux supplémentaires autour de cette notion, de manière à pouvoir montrer que c'est, en volume, la thématique la plus souvent citée. Le travail de découpage et

de regroupement des codes est donc soumis à un certain arbitraire d'un point de vue purement linguistique et sémantique, mais qui se justifie par le contexte et la fonction de l'analyse textuelle.

La négociation de la façon dont le plan de code est adapté et alimenté met en évidence la complexité des normes épistémologiques à l'œuvre au sein d'un même projet. Tout d'abord, il y a trois niveaux de normes : celles qui régissent l'identification des éléments lexicaux intéressants (les annotations), la constitution du plan de code, et l'articulation entre les deux matérialisée par la façon dont le plan de code est alimenté par des éléments lexicaux.

La première étape est la plus consensuelle. C'est la moins accessible pour le client qui se fie aux outils de Proxem et aux compétences des infolinguistes pour faire ce travail fastidieux. Du point de vue du client, c'est une épistémologie de la confiance qui régit la validité des résultats de cette tâche. Cette confiance vient notamment du fait que la matérialité du texte est perçue de manière non problématique : c'est un objet dont la positivité permet de dégager mécaniquement, objectivement (pour le client), des éléments « naturellement » pertinents. Du point de vue de l'infolinguiste, la tâche est un peu plus problématique même si l'intérêt de certains éléments lexicaux a un caractère d'évidence. Le fait même que certains éléments puissent ne pas être intéressants est bien connu de l'infolinguiste mais peut être une surprise pour le client qui aimerait « tout » prendre en compte et évalue parfois le travail de l'infolinguiste en quantité ou proportion de termes annotés ; pour l'infolinguiste au contraire, des segments parfois importants des verbatims sont d'un intérêt limité dans une perspective de codification. Par ailleurs, il y a une distance, souvent inconnue du client, entre ce qui existe dans le corpus, et ce qui est accessible *via* la syntaxe des règles linguistiques : les infolinguistes savent que la détection peut être imparfaite, et en particulier que l'ambiguïté d'un terme peut être évidente pour un lecteur humain, mais ne peut pas, ou difficilement, être capturée par des règles formelles. Cette difficulté relève à la fois d'une limitation pratique de l'outil, mais aussi (et peut-être surtout dans certains cas) d'une limitation théorique. Ce sont les éléments problématiques qui remettent en cause la positivité du sens des éléments pour revenir à une épistémologie herméneutique, où le sens émerge par la lecture interprétative et n'est plus contenu de manière permanente dans la matérialité du texte. Dans ce cas le moteur d'analyse sémantique n'est pas capable d'une lecture interprétative, mais instrumente celle des infolinguistes. Cela reconfigure le rôle des objets techniques, qui ne sont pas producteurs de connaissance, mais support et instrument de l'activité épistémique. Néanmoins, que la posture des infolinguistes soit réaliste ou herméneutique n'a guère d'influence sur le travail réalisé, qui dépend davantage de modalités pratiques (techniques) que de positions théoriques.

Les phases de constitution et d'alimentation du plan de code sont plus problématiques. Comme on l'a vu, en l'absence d'un plan de code idéal dont on se rapprocherait, il lui faut une autre légitimation que son caractère « naturel ». Le plan de code n'aura pas un degré de perfection mais un certain nombre de vertus épistémiques qui contribueront à le stabiliser. Parmi ces vertus, on compte :

- son empiricité (il y a effectivement des verbatims correspondant aux codes) ;
- son homogénéité (le nombre de verbatims assigné à chaque code relève du même ordre de grandeur) ;
- sa cohérence (les codes se complètent et ne se superposent pas) ;
- sa complétude (toutes les thématiques abordées dans le corpus sont couvertes) ;
- sa lisibilité (le libellé du code permet de comprendre comment il est alimenté) ;
- son utilisabilité (il permet au client de remplir ses objectifs).

Bien que les premiers points soient valorisés, c'est en pratique le dernier qui gouverne l'ensemble. Peu importe en fin de compte que le plan de code soit incohérent ou non représentatif des verbatims, pourvu qu'il permette au client et à l'entreprise de remplir ses objectifs. Ses thématiques doivent refléter des indicateurs intelligibles qui lui permettent de mesurer et de décider. De ce fait, on passe d'une posture où le plan de code doit refléter la réalité pour être utilisable, à une posture différente, où il est valide dès lors qu'il est utilisable. La validation est marquée par la fin de l'intervention de l'infolinguiste sur le projet, et non sur la finalisation d'un plan de code idéal. Aussi, tout en formulant ses remarques à l'infolinguiste en des termes réalistes, le client apporte en réalité, comme on le verra plus en détail au chapitre IX, une épistémologie pragmatiste, où la connaissance est ce qui permet l'action.

Ce pragmatisme régit par rétroaction l'ensemble des étapes. Les infolinguistes ne visent pas par leur travail ce qui leur semble scientifiquement valide suivant des normes linguistiques, mais ce qui est demandé par le client. Elles adoptent donc une posture relativiste, ouverte au pluralisme des interprétations et au caractère contestable de l'analyse. La négociation autour du redécoupage et regroupement des codes accentue ce relativisme : si le premier plan de code qu'elles proposent au client peut leur sembler présenter beaucoup de vertus épistémiques, les multiples négociations de cette première proposition conduisent à un résultat final qui peut sembler particulièrement arbitraire. Cet arbitraire n'est pas quelque chose qui relève de l'aléatoire ou du non-sens, mais qui ne possède en revanche aucun caractère nécessaire.

En synthèse, la codification est une activité prise en charge opérationnellement par les infolinguistes, mais qui procède d'une épistémologie sociale : ses normes de constitution et de stabilisation sont réparties entre des agents distincts qui ont des visées différentes. Chaque agent traverse un certain nombre de postures épistémiques qui ne sont, pour leur part, pas stables :

- les infolinguistes peuvent adopter initialement une posture plutôt réaliste en ce qui concerne la détection d'entités nommées, où la mise en évidence d'éléments thématiques a un statut d'assertion (c'est-à-dire qu'elle peut être vraie ou fausse), et peut donc être corrigée – ou, dans des termes plus poppériens, falsifiée. Néanmoins, la diversité des sources de ressources linguistiques, ainsi que la rencontre de cas limites qui mettent le logiciel en difficulté,

ramènent à une posture plus herméneutique, où le sens émerge de la lecture et rend possible un pluralisme des interprétations. La posture par rapport au plan de code peut évoluer, d'un mélange d'empirisme et de rationalisme cohérentiste, à un nominalisme pragmatiste ;

– le client démarre avec une posture réaliste en ce qui concerne les codes mais prend progressivement conscience de leur arbitraire. Il demeure généralement réaliste en ce qui concerne la détection d'entités nommées, qui peut (et doit) être corrigée, ce qui veut dire qu'elle s'inscrit dans une épistémologie où elle peut être bonne (juste) ou mauvaise (fausse) ; les entités nommées correctement détectées ont un statut de connaissances vraies. En revanche, sa posture par rapport au plan de code évolue pour lui aussi vers un pragmatisme qui régit l'ensemble, tout en maintenant un certain nombre de vertus épistémiques comme la cohérence ou l'empiricité, mais qui n'ont pas un statut de normes obligatoires.

8. La mise en visibilité de l'activité linguistique

Le plan de code finalisé n'est pas le seul élément qui reçoit une forme de restitution. Le travail de négociation qui aboutit au plan de code est lui aussi matérialisé, par des objets intermédiaires qui complètent les échanges oraux ou écrits entre l'infolinguiste et le client. De nombreux fichiers Excel circulent, parmi lesquels certains sont relativement standardisés. L'un d'eux permet de centraliser les remontées faites par le client, qu'il s'agisse d'erreur de détection ou de demande de réorganisation du plan de code. L'organisation en colonnes permet de tracer les demandes mais aussi leur exécution, les éventuelles remarques de l'infolinguiste, etc.

Néanmoins, la principale forme graphique est le logiciel développé par Proxem et mis à disposition des clients (voir **Figure 12**). Jusqu'en 2017, celui ne sert pas à configurer l'analyse sémantique mais seulement à la restituer au client. L'intérêt de la forme logicielle par rapport à un support statique (comme une page web basique, un rapport, un document imprimé) est d'intégrer des éléments d'interactivité. Il présente pour cela une interface qui donne à voir autant qu'elle permet de faire. Concrètement, le ou les plans de code et les métadonnées sont matérialisés sous la forme de volets appelés « facettes », selon une terminologie issue des moteurs de recherche. À côté de chaque code est indiqué, entre parenthèses, le nombre de verbatims correspondants. Initialement, le logiciel était en effet conçu comme un moteur de recherche à facettes, permettant plusieurs sortes de filtres et de consulter les verbatims correspondant à ces filtres. Un bouton permet de matérialiser les entités nommées détectées dans les verbatims, en les soulignant. Les codes listés dans chaque facette sont cliquables : le clic permet de filtrer sur le sous-ensemble du corpus correspondant au code. Les codes sont aussi accessibles par autocomplétion *via* une barre de recherche qui permet également une recherche « plein texte » de mots ou expressions. Depuis 2012, des recherches dites « avancées » peuvent être faites en combinant des codes et des expressions avec des opérateurs booléens « AND »,

« OR » et « NOT ». Lorsqu'une recherche est effectuée, les comptages correspondant aux codes sont mis à jour dynamiquement pour correspondre à ceux de la sélection.

Aux fonctionnalités de recherche se sont progressivement ajoutées des possibilités de visualisation. Une vue « statistique » permet de visualiser les comptages correspondant à une facette sous la forme d'un diagramme en bâton ou d'une vue arborescente *(treemap)*. Lorsque les verbatims sont datés, un onglet représente l'évolution du volume dans le temps sous la forme d'un diagramme linéaire. Un nuage de termes *(tagcloud)* peut être généré sur la base d'une extraction terminologique ou d'autres méthodes d'extraction d'information. Enfin, une carte de chaleur *(heatmap)* permet de croiser deux facettes pour visualiser des surreprésentations statistiques.

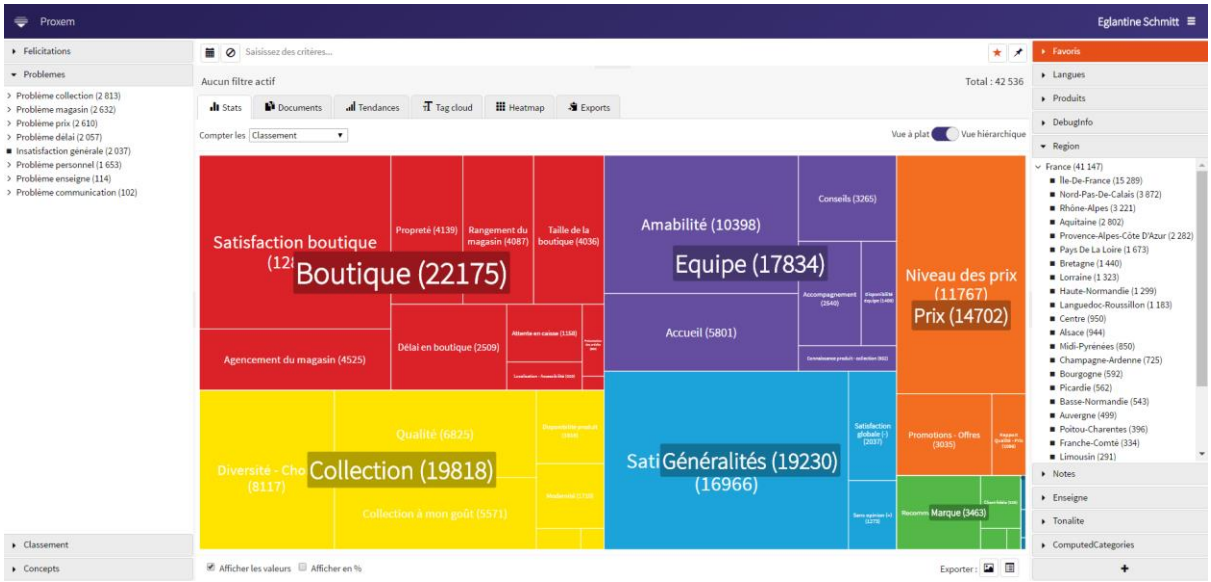

***Figure 12**. Capture d'écran du logiciel montrant une vue hiérarchique du plan de code (volet central), le plan de code intermédiaire* Problèmes *(volet latéral gauche) et la métadonnée* Région *(volet droit).*

À travers ces fonctionnalités de visualisations, fondées sur des comptages, l'outil de recherche devient également un outil d'analyse statistique. Il emprunte des conventions d'interface, d'interaction et de représentation issus de ces différents univers. Ces fonctionnalités combinées aux usages visés inscrit le logiciel dans l'héritage de l'informatique décisionnelle *(business intelligence)* et de ses outils fondés sur l'exploration statistique de données dites *métier* pour la prise de décision en entreprise.

Le logiciel et son interface sont le résultat d'un travail de conception qui relève de l'ingénierie et du design, et sur lequel nous reviendrons. Retenons pour le moment que le logiciel évolue au quotidien en parallèle des projets menés par Proxem. Cette évolution s'inspire des remarques des clients et de l'analyse de leurs usages mais aussi de pistes identifiées en interne. En particulier, Proxem dispose d'un pôle « Études » que j'ai constitué et qui est un utilisateur intensif du logiciel. Plutôt que de restituer le travail des infolinguistes à travers le logiciel, ce pôle propose en effet aux clients de Proxem des rapports d'étude qui rendent compte d'une analyse à la fois linguistique et statistique. Au fil des

projets, une méthode de travail s'est constituée pour réaliser ces études ; cette méthode s'est enrichie avec l'évolution du logiciel et a aussi été à l'origine de nouvelles fonctionnalités. Concrètement, les fonctions d'export et de visualisation de données ont été ajoutées sous l'impulsion des besoins des chargés d'étude, et en cohérence avec les besoins des clients de Proxem utilisateurs du logiciel.

9. La mise en histoire et l'interprétation des données

Cette méthode joue ainsi un rôle épistémologique déterminant dans l'intelligibilité des données dont rendent compte les rapports d'étude, mais aussi dans les usages des clients. Comme tout objet technique, le logiciel prescrit des usages en les rendant possibles ; plus il se rapproche d'une implémentation de la méthode de travail du pôle études et plus c'est cette méthode qui est prescrite aux clients à travers le logiciel, indépendamment de toute formation à l'outil qui expliciterait cette méthode. Cette méthode devient le modèle d'usage par défaut du logiciel, en particulier pour les clients de Proxem qui n'ont pas un cadre d'utilisation fortement prédéfini, par exemple parce que le projet mené avec Proxem a pour eux un statut prospectif d'expérimentation.

Cette méthode a également un statut épistémologique particulier : elle n'est pas purement empirique ou heuristique, ni le pur produit d'une tradition disciplinaire. Elle est partiellement formalisée et transmise surtout oralement et par l'exemple. Elle puise dans des héritages scientifiques et techniques (et en particulier la tradition française de statistique textuelle appliquée aux sciences sociales) tout en expérimentant à partir de ce que le matériau brut dicte comme possibilités. Enfin, elle occupe un statut spécifique par rapport à la thèse présentée dans ces lignes : elle est largement nourrie par le travail d'observation et de recherche de ladite thèse. Si la présentation de la démarche de codification relevait jusqu'ici d'une approche ethnographique, l'explicitation de la méthode de travail du pôle étude est un travail réflexif de description et de conceptualisation.

Le socle de cette méthode est une approche quantitative, très largement inspirée des statistiques pour les sciences sociales. En effet, la codification des verbatims est un processus de quantification qui a pour résultat d'assigner un poids chiffré aux différents codes, ce poids correspondant au nombre de verbatims attribués à chaque code. La quantité est d'abord descriptive : elle rend compte des données textuelles de manière agrégée en indiquant de quoi parlent les verbatims et dans quelles proportions. À la lecture linéaire de chaque verbatim, elle substitue une lecture synthétique, dont la forme graphique va être celle de la visualisation de données.

La première visualisation d'une étude consiste en effet à rendre compte du plan de code sous la forme d'un diagramme en bâton, ou d'une vue arborescente de type *treemap* pour les plans de code à deux niveaux. À ce titre, elle s'apparente aux visualisations utilisées par les instituts de sondage pour rendre compte de la distribution des réponses à une question fermée. Néanmoins, contrairement à une question comme l'âge du répondant, le plan de code n'est pas une partition, c'est-à-dire un découpage

où chaque élément est assigné à une catégorie et une seule. D'une part, plusieurs codes peuvent être assignés à un verbatim et d'autre part, certains verbatims n'ont aucun code. De ce fait, le plan de code quantifié ne peut pas être considéré comme une répartition, et ne doit pas être représenté par un diagramme circulaire (dit « camembert »). Tout d'abord, comme on le reverra au chapitre VIII, ce type de visualisation est sévèrement décrié depuis plus de 30 ans (Cleveland et McGill 1985) par la communauté de statisticiens, psychologues et designers qui s'intéresse à ces problématiques (Su 2008), à commencer par Tufte (1997) que nous avons déjà présenté. Cette communauté s'appuie notamment sur des recherches empiriques évaluant la lisibilité des différents types de visualisation en comparant le temps nécessaire pour déterminer quelle est la plus grande mesure, la plus petite, sur différents graphiques. La représentation de l'information quantitative par des angles rend la comparaison entre mesures particulièrement difficile, notamment lorsque le diagramme circulaire est déformé par un effet 3D (Few 2007). Par ailleurs, le diagramme circulaire ne peut être utilisé que pour représenter une série de données qui constitue une partition, c'est-à-dire qui peut être décrite par une série de pourcentages dont la somme fait 100%, ce qui n'est jamais le cas des éléments d'un plan de code.

De plus, si la mesure de la distribution des âges au sein d'une population peut être dégradée par un mauvais échantillonnage, un mauvais taux de réponse, des répondants mentant sur leur âge, la quantification des codes est également sensible aux erreurs de codification. Aussi, quand bien même tous les verbatims seraient classés, avec un seul code, les erreurs dans cette répartition seraient de nature différente des erreurs présentes dans des réponses à des questions fermées. En cela également la codification se distingue du travail des instituts de sondage et suggère des normes de validité différentes. Il faut souligner par ailleurs que le taux d'erreur sur les codes tend à décroître avec le volume : plus il est important et plus la hiérarchisation des codes est fiable car le nombre d'entités nommées correctement détectable augmente avec le volume. De ce fait, la quantification des codes s'inscrit dans une raison statistique où la fiabilité des résultats converge selon la loi des grands nombres (Desrosières 2010) ; en cela, la précision de la codification augmente avec le volume et non avec la qualité de l'échantillonnage.

La quantification des codes est une forme de mise en évidence de régularités, d'une norme statistique. Elle a un caractère nomologique, qui révèle une forme de constance. Néanmoins, elle peut également servir de base à une analyse qui va au contraire rechercher les anomalies, les irrégularités, les éléments surprenants. Dans cette perspective, le plan de code quantifié est le point de départ d'une analyse exploratoire qui s'apparente à la démarche décrite par John Tukey et appliquée en fouille de données. La recherche d'aspérités par rapport au plan de code se fait en le croisant avec d'autres angles de l'analyse : l'information temporelle ou géographique, les métadonnées, mais aussi les autres plans de code. Ce croisement prend une forme connue de la statistique multivariée : le tableau de contingence, qui permet traditionnellement d'estimer la dépendance entre deux variables. Néanmoins, dans le

contexte de Proxem le tableau croisé n'est pas utilisé pour produire une synthèse de l'information sous la forme d'un score de dépendance. Les scores classiques, comme celui produit par le test du χ^2, sont utilisés et considérés comme valides pour des échantillons de l'ordre du millier d'éléments, ou des volumes inférieurs. Sur des volumes importants, ce score, corrélé au volume, finit par n'exprimer rien d'autre que la taille de l'échantillon (Gelman et Shalizi 2013). Empiriquement, j'ai constaté que le score de dépendance entre le plan de code et un autre angle d'analyse est toujours statistiquement significatif selon les normes de la statistique classique. Ce score n'est donc pas exploitable en tant que tel ; en revanche, la variabilité d'un angle d'analyse par rapport à un autre, capturable par exemple en calculant le résidu local du χ^2 pour chaque comptage du tableau de contingence, permet de détecter des aspérités, favorisant des analyses de nature idiographique.

Du fait de l'inapplicabilité des tests d'hypothèse, on passe ainsi d'une analyse nomologique productrice de normes, à une analyse idiographique qui recherche de la singularité à travers de la variabilité. Les croisements statistiques dessinent un espace à décrire et à explorer comme on le ferait d'un espace géographique. Cette phase exploratoire s'apparente ainsi à l'approche développée par Marc Reinert dans le cadre du logiciel d'analyse de données textuelles Alceste, et qu'il décrit également en des termes cartographiques :

> notre démarche ressemble davantage à la démarche d'un cartographe, qu'à celle d'un chercheur d'or. Il s'agit d'abord d'explorer un monde inconnu dans ses principaux reliefs, avant de tenter de s'y frayer un chemin, en fonction de ses intérêts, en fonction aussi des aléas du terrain, pour trouver l'or du sens. (Reinert 1990)

Cette perspective se rapproche ainsi aisément des approches statistiques multivariées fondées sur l'analyse de la variance, et tout particulièrement des analyses factorielles de l'école française de statistique comme l'analyse de correspondance mise au point par Jean-Paul Benzécri dans une perspective propice à l'analyse de questions ouvertes. Comme le rappelle en effet Michel Armatte (2008), le travail de Benzécri dans les années 1960,

> centré sur une approche inductive de la linguistique par l'analyse de tables de distributions de mots (dans la lignée des travaux de Z.S. Harris, opposée à celle de Chomsky), forme dès 1962 le cadre de l'invention de l'analyse des correspondances comme méthode d'analyse factorielle de tableaux de contingence munis d'une métrique du χ^2 imposée par le principe d'équivalence distributionnelle.

Il est remarquable en effet que l'analyse factorielle de correspondance, qui sera par la suite utilisée pour toutes sortes de données catégorielles structurées, soit initialement imaginée dans une perspective linguistique inductiviste, fortement compatible dans son origine avec la démarche développée par Proxem aujourd'hui. La méthodologie d'Alceste s'en revendique d'ailleurs également. Les croisements entre le plan de code (plutôt que la distribution de mots) et une autre variable sont matérialisés dans le logiciel par une carte de chaleur interactive *(heatmap)* des écarts à l'indépendance,

conçue dans la perspective du pôle études, mais on les retrouve également sous forme de plans factoriels dans les rapports d'étude proposés par Proxem. En croisant le plan de code avec l'ensemble des questions fermées d'une enquête, on obtient un tableau de Burt similaire au « tableau lexical des questions » développé par le sociologue Philippe Cibois (1990) qui s'appuie, comme chez Benzécri, sur la distribution du vocabulaire. Ce tableau lexical des questions est soumis à une analyse factorielle de correspondance qui met en évidence les attractions entre mots et caractéristiques sociodémographiques ou opinions **(Figure 13)**.

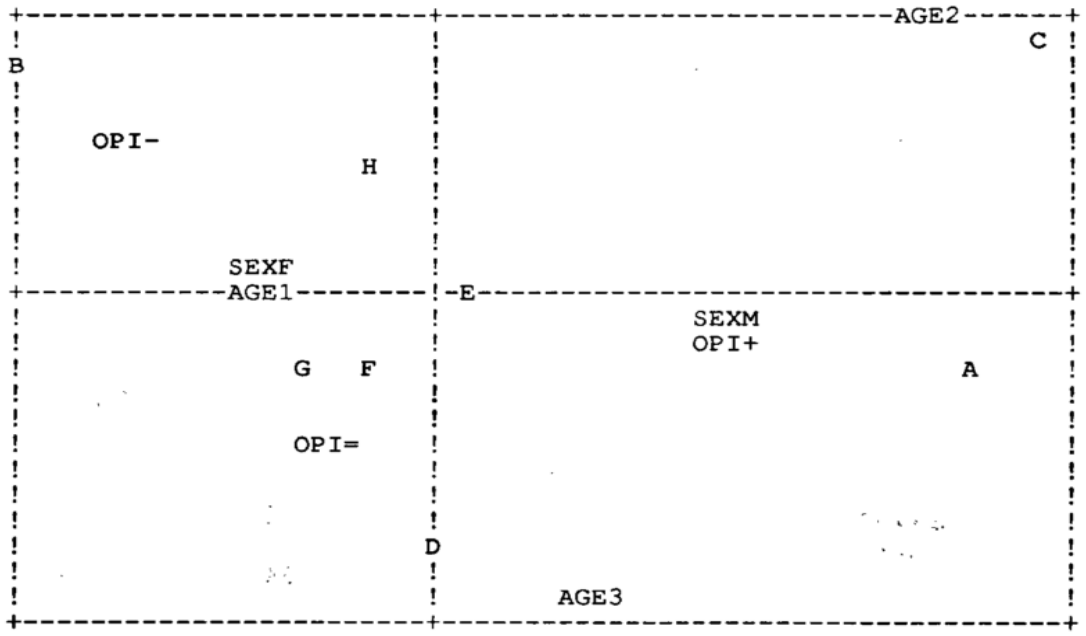

***Figure 13**. Exemple de plan factoriel mettant en relation le vocabulaire des réponses à une question ouverte (représenté par des lettres A, B, C, etc.) et les réponses aux questions fermées (sexe, âge et opinion). Tiré de Cibois, 1990.*

Une autre manière d'envisager ces croisements est sur le mode de la cooccurrence. Lorsque l'on croise deux axes issus de l'analyse sémantique (par exemple deux plans de code), voire un plan de code avec lui-même, cela revient à calculer la matrice de cooccurrence des codes, qui correspondent à des occurrences linguistiques de termes. Or la cooccurrence linguistique a elle aussi une histoire méthodologique et une tradition d'interprétation ; c'est une notion déjà chargée de sens, au-delà de sa manipulabilité computationnelle. En linguistique distributionnelle, les termes qui ont les mêmes voisins sont considérés comme similaires, voire synonymes. Le voisinage d'un terme est ce qui permet d'en déterminer le sens : dans la sémantique interprétative de François Rastier (2001; 2015) par exemple, le sens d'un terme est conceptualisé comme une différence par rapport à son contexte, et le sens d'un texte dépend du corpus dont il fait partie. Comme le résume la rastierienne Bénédicte Pincemin (2012), les cooccurrences fréquentes, ou statistiquement significatives, peuvent être promues en corrélations au sein d'un parcours interprétatif qui en valide la signification :

> Ce que produit le calcul textométrique, ce sont donc des cooccurrents, au plan des signifiants ; et ce qui est visé, c'est l'obtention de corrélats, au plan des signifiés. On passe des premiers aux seconds par une interprétation qui reconnaît la présence d'un trait sémantique commun entre le ou les mots servant d'amorce à la recherche, et le cooccurrent alors qualifiable comme corrélat.

Cette analyse de cooccurrences présente deux caractéristiques : la première, héritée de la tradition textométrique, est que la cooccurrence ne révèle rien par elle-même, mais doit être interprétée pour être validée et promue en corrélation (voire éventuellement en rapport de causalité) *via* un parcours interprétatif. La seconde est que l'analyse n'est jamais décrite telle quelle, sous la forme d'un texte seul, d'une écriture linéarisée, mais s'accompagne systématiquement d'une visualisation : carte de chaleur, plan factoriel, ou encore visualisation de graphe dans un logiciel spécialisé indépendant tel que Gephi (Bastian, Heymann et Jacomy 2009). Pour cette troisième option, les codes sont représentés par des nœuds et les cooccurrences par des liens, en cohérence avec les habitudes d'analyse de réseaux thématiques ou de cooccurrence lexicales en pratique depuis plus de 20 ans (Grefenstette 1994; Leydesdorff 1995).

Dans ces trois options, la visualisation joue un double rôle d'objet intermédiaire, heuristique, et de forme de restitution de l'étude ; dans les termes de Reichenbach, elle s'emploie dans un contexte de découverte comme dans un contexte de justification. Comme nous le soulignions plus haut, les croisements statistiques tracent un espace à décrire : le rôle de ces visualisations intermédiaires est alors précisément de cartographier cet espace à explorer *via* sa représentation, et à interpréter pour en faire ressortir, toujours dans une perspective idiographique, les indices pertinents. Dans la forme de restitution, ces indices sont mis en évidence par des signes graphiques qui reconstituent l'interprétation : zones entourées (ou flèches) accompagnées d'un commentaire qui fournit au lecteur une représentation déjà interprétée, tout en laissant la visualisation, toujours visible, jouer son rôle de preuve quantitative et topologique.

Cette rhétorique visuelle, sur laquelle nous reviendrons au chapitre VIII, s'accompagne d'une rhétorique narrative, celle du rapport d'étude lui-même. Constitué de diapositives PowerPoint, il n'est une simple juxtaposition paratactique de visualisations de données sans rapport entre elles. Il est conçu comme un récit qui présente une progression, des étapes et une cohérence. La structure de ce récit présente des invariants d'un rapport à l'autre, mais chaque étude est unique du point de vue de son contenu. Ce caractère narratif et argumentatif est renforcé par la pratique systématique d'une soutenance orale où la chargée d'étude[41] défend sa démarche et son point de vue sur les données. Le rapport a ainsi un double statut de reconstitution de l'investigation par son récit, et de restitution raisonnée conduisant sur un mode argumentatif d'hypothèses vers une conclusion. Le temps de la soutenance orale est aussi un temps de discussion avec le client, qui interrompt le récit, demande des précisions ou des clarifications, complète des pistes, commente les hypothèses, etc. Le commentaire

[41] Le pôle étude lui aussi n'a connu jusqu'à ce jour que des profils féminins.

oral favorise la compréhension en explicitant des choix graphiques ; il peut aussi intervenir à un niveau de commentaire supplémentaire, en portant un jugement sur les commentaires écrits dans le rapport. C'est l'ensemble constitué par la représentation graphique des visualisations, la forme narrative, et enfin le temps du récit et du dialogue oraux, qui construit l'intelligibilité des données pour le client, qui peut alors se prononcer sur la validité des connaissances produites, et partager à son tour l'étude réalisée avec d'autres intervenants. Qu'il s'agisse de la codification des verbatims eux-mêmes, ou de l'analyse statistique des résultats, le sens du corpus n'en émerge pas purement comme un calcul, mais au terme d'un parcours interprétatif fondé sur les données textuelles, où se succèdent des validations sémantiques, statistiques et pragmatiques par des acteurs différents travaillant sous des régimes épistémiques qui leur sont propres.

Conclusion

Nous avons présenté deux modes d'analyse computationnelle de données textuelles en entreprise et relevant des sciences de la culture à partir du cas concret de l'entreprise Proxem. Ces deux modes, celui du logiciel et celui de l'étude, ne sont pas exclusifs mais complémentaires :

- D'une part, le **logiciel**, comme outil informatique, joue le rôle de transcendantal empirique que nous avons mis en évidence au chapitre V. Il est en effet la condition de possibilité matérielle de l'étude.
- D'autre part, la **méthode** employée par cette dernière est ce qui guide le développement et les évolutions du logiciel. Celui-ci constitue l'implémentation matérielle des normes, gestes, et successions d'étapes propres à cette méthode.

Ainsi le logiciel n'automatise pas le traitement de données textuelles, mais programme des actions ; il réduit le champ des possibles non pas jusqu'à le ramener à un unique déroulement nécessaire, mais en l'inscrivant dans un modèle d'usage pensé par la méthode d'analyse. Celle-ci s'inscrit dans plusieurs lignées méthodologiques issues à la fois de la linguistique, des méthodes quantitatives en sciences sociales, et des rencontres qui se sont produites entre les deux, dont notamment celle qu'incarne l'école française de statistique. L'articulation entre logiciel et méthode est similaire à celle d'un autre outil marquant de cette histoire, le logiciel Alceste développé par Marc Reinert dans les années 1980 (Reinert 1990), et dont l'implémentation dans le logiciel libre Iramuteq (comme surcouche du langage de programmation R) par Pierre Ratinaud et Sébastien Déjean (2009) constitue aujourd'hui une référence dans le monde universitaire pour l'analyse de contenus textuels.

Par rapport à cette tradition méthodologique, la démarche adoptée par Proxem présente un certain nombre de ruptures. Tout d'abord, elle ne s'en revendique pas explicitement : c'est le travail généalogique que j'ai mené qui en a révélé les similarités et a apporté la culture épistémique propre à la codification et à la statistique pour les sciences sociales. Par ailleurs, la nature de la donnée induit

un certain nombre de ruptures fortes par rapport à l'orthodoxie des méthodes quantitatives en sciences sociales. En effet, l'agrégation de données issues de sources multiples, suivant des logiques hétérogènes, abolit la norme de la représentativité, centrale en statistique, pour la remplacer par la compatibilité des contextes pragmatiques des différentes sources, commensurable sous l'angle des attentes du client de Proxem.

De plus, la division du travail induit une rupture forte entre la constitution et l'analyse des données. Pour les infolinguistes et chargées d'étude, la donnée est un donné envisagé dans sa positivité et non dans sa qualité de représentation qui fait signe vers un certain réel. C'est aussi un objet matériel qui doit être refondu dans un certain format technique. Son statut de représentation du réel ne pourra être reconstruit qu'au terme de l'analyse, par l'interprétation rendue par la visualisation et la narration. Avant cette restitution, les phases de nettoyage de données, les choix de gestion des données manquantes, ou à la granularité hétérogène, ainsi que les phases d'exploration et de cartographie, rappellent quant à elles davantage la culture épistémique de l'analyse exploratoire de données à la Tukey et de la découverte de connaissances dans les bases de données.

La démarche globale procède ainsi d'une hybridation des héritages épistémiques de l'analyse de données textuelles en sciences sociales et de l'informatique d'entreprise *(business intelligence)*. Cette hybridation se matérialise également dans une épistémologie sociale où les cultures épistémiques des différents agents conditionnent leurs apports respectifs au résultat final. Elle se nourrit de ces héritages, adopte certaines de leurs vertus épistémiques, tout en étant d'une part tributaire d'un savoir-faire, d'une *technè* difficilement formalisable, et d'autre part régie par une épistémologie pragmatiste, à la fois dans son sens courant de manière de faire qui s'adapte aux contraintes sans allégeance forte à des principes théoriques généraux, et dans son sens philosophique de conception de la connaissance déterminée par ses conséquences pratiques.

Chapitre VII.
Du récit au système : les formes de restitution de l'analyse de données

Dans les précédents chapitres, nous avons présenté deux composantes de notre paradigme méthodologique nécessaires pour la production de connaissances à travers l'analyse computationnelle de données massives : la capacité à conférer une valeur épistémique aux *big data* à travers leur statut indiciaire, et le rôle de transcendantal empirique, autonome par rapport aux cultures épistémiques dans lesquelles il est mobilisé, que jouent les traitements computationnels des sciences des données. Dans ces deux composantes, la dimension idiographique et exploratoire revient comme un *leitmotiv* qui contredit la vision largement répandue des *big data* comme processus de quantification et d'automatisation par la donnée et le calcul. Notre étude de cas a ensuite mis en évidence que ces composantes sont en effet nécessaires, mais pas suffisantes, pour dégager une intelligibilité des données : elles sont complétées par un savoir-faire lui-même guidé par le contexte pragmatique du projet épistémique dans lequel intervient l'analyse de données, et qui varie en fonction de la nature et de la valeur des données.

Plus spécifiquement, notre étude de cas présentait un contexte pragmatique stabilisé : celui d'un client qui sollicite un prestataire pour un projet de codification sur la thématique de l'expérience client. Cette stabilité permet la formation d'habitudes qui se sédimentent pour constituer progressivement une méthode, itérativement intégrée dans un logiciel dédié. Dans les termes de Simondon (1958) parlant des machines, le logiciel est « du geste humain fixé et cristallisé en structures qui fonctionnent ». Cette étude de cas illustre ainsi une dynamique de transformation entre les deux configurations que peut avoir un projet de traitement de données massives, et que nous présentons maintenant ici : la **configuration en récit** et la **configuration en système**. Ces deux configurations se distinguent principalement par leur forme de restitution, mais aussi par le processus d'analyse de données qui permet d'aboutir à cette restitution. Elles constituent les deux extrémités du spectre dans lequel s'inscrivent les projets d'analyse de données massives.

2.1. Deux configurations épistémiques

La **configuration en récit** suit ainsi un processus classique de production de connaissances : une problématique est posée, on recherche des éléments empiriques permettant d'y répondre, puis on livre les résultats sous la forme d'un récit oral ou écrit, typiquement un rapport, une conférence ou une publication scientifique. Ce récit est le livrable d'un projet délimité dans le temps, dont il présente les résultats sous une forme écrite qui retrace plus ou moins linéairement l'investigation, mais dont la forme d'expression est linéaire. Le récit reconfigure le temps de l'investigation et lui donne cette forme linéaire.

Il s'agit donc bien d'un récit qui réalise une synthèse temporelle de l'hétérogène (Ricœur 1983), et s'inscrit dans un temps de lecture ou de diction, distinct du temps de l'événement et du temps de l'enquête. Le récit est une forme narrative qui peut mobiliser toute la diversité des leviers rhétoriques et narratifs visant à obtenir l'intérêt, l'attention, l'approbation ou encore la confiance du lecteur ou de l'auditoire. La légitimité du récit ne vient pas tant du régime de preuve qu'il peut instaurer que de sa cohérence narrative et de sa capacité à entraîner le lecteur, à le faire entrer dans cette cohérence. Le pacte qui s'instaure entre le narrateur et le lecteur dans le cas d'une œuvre de fiction, pour laquelle ce dernier accepte d'entrer dans un autre référentiel que celui du monde qu'il connaît, s'instaure également ici pour autoriser le narrateur à rendre manifeste et intelligible ce qui n'est pas là, comme dans le récit historique de Ricœur. Les leviers du roman policier se prêtent particulièrement bien à cette configuration : il s'agit, en effet, de rendre compte d'une investigation, ce que l'on peut faire en suivant la chronologie de l'investigation, mais aussi celle du « crime », ou encore déconstruire cette linéarité en commençant par dévoiler toutes les conclusions, puis en révélant comment l'investigation a permis d'y parvenir.

Du point de vue du processus de l'enquête (et non du récit), les éléments empiriques visant à répondre à la problématique sont des données numériques qui sont plus logiquement recueillies après la formulation de la problématique, même s'il peut arriver, à l'inverse, que leur obtention déclenche le projet épistémique lui-même. Dans le cas où les données sont préexistantes, le projet sera alors en substance « de quoi parlent ces données et que puis-je en faire ? ». Ainsi, parmi les clients de Proxem par exemple, tous n'ont pas un besoin spécifique adossé à un métier de l'entreprise. Le phénomène *big data* a propagé en entreprise une forme d'injonction à lancer des projets d'analyse de données numériques sans objectif précis, initiés soit par des départements « innovation » existants, soit par des créations de poste comme celle d'un « chief data officer ». Ces intervenants ont généralement pour mission d'évaluer, d'expérimenter, autour des outils et des usages de la donnée ; de ce fait, le traitement de données est pour eux une fin en soi, qui n'est pas adossée à un projet épistémique, telle qu'une hypothèse à confirmer ou un phénomène à mesurer. Ces projets sont les plus difficiles pour Proxem, car ils n'ont pas de critères de succès. Ils tendent cependant à se raréfier à mesure que le

phénomène *big data* arrive à maturation dans les entreprises, lesquelles parviennent à définir des usages de la données adossés à ses métiers.

Par ailleurs, la configuration en récit est aussi présente dans la production scientifique en informatique, en *machine learning*, et dans les sciences de la culture mobilisant des données numériques comme les *digital methods* : la restitution sous forme d'article de recherche reste la norme du monde universitaire, même si l'effort de captation du lecteur et de mise en tension du lecteur y est moins forte que dans d'autres contextes. Dans notre cas, il s'agit d'articles qui relatent l'exploitation computationnelle d'un jeu de données caractérisées comme « *big data* ». Enfin, cette configuration en récit est aussi la forme privilégiée d'usages émergents de la donnée, comme le journalisme de données, dont le livrable est généralement un article agrémenté de visualisations de données, ou encore le *data storytelling*, dont le nom explicite sa forme de récit tout en servant des usages variés, amateurs, professionnels, universitaires ou autres. La notion de *data storytelling* pourrait d'ailleurs désigner toute forme de restitution issue de la configuration en récit.

Dans tous les cas, la configuration en récit relève de l'artisanat et de l'innovation : il s'agit de retracer par le récit une articulation nouvelle entre un projet épistémique et des données analysées, entre des hypothèses ou intentions de connaissances, et une manière de les tester. Le projet épistémique peut être ancien, mais n'avoir jamais exploité le type spécifique de données utilisé en l'occurrence : l'enjeu est alors de retrouver *via* l'analyse de traces numériques des connaissances produites autrement par le passé. À l'inverse les données permettent de faire émerger un nouveau projet épistémique, voire un nouveau genre de projet épistémique : lorsqu'un *data scientist* se voit confier un jeu de données pour les analyser, il n'y a pas d'hypothèse prédéfinie à confirmer ; il lui faut imaginer les connaissances qu'un tel jeu de données pourrait faire émerger, et ce dans le cas où c'est la première fois que ces données sont étudiées, ou non. Il n'est pas nécessaire que le projet épistémique ou que les données soient inédits, mais que l'articulation entre les deux le soit.

De ce fait, ce caractère innovant donne à ce type de configuration un tour presque nécessairement artisanal, dans la mesure où une articulation inédite suppose non seulement un savoir-faire, c'est-à-dire la capacité à mettre en œuvre une façon de faire, mais également de l'inventivité dans la méthode et la procédure suivie, c'est-à-dire la capacité à imaginer de nouvelles façons de faire. Elle implique qu'il n'existe pas de procédure formelle stabilisée préalablement au projet épistémique, car la mise au point de cette procédure est précisément le produit du projet dont le récit rend compte ; c'est pourquoi on parle d'invention. Appliquer la méthodologie et les normes d'une science spécifique n'est pas une option car l'articulation en jeu relève par essence d'une innovation méthodologique qui déborde des normes épistémiques existantes ; le récit rend compte de la solution inventée pour un problème inédit. Cette solution est constituée d'actions sur les données, d'outils développés *ad hoc* ou réassignés, et d'une logique interprétative. Par exemple, dans les rapports proposés par Proxem, la mobilisation de l'analyse factorielle de correspondance développée par Benzécri intègre la mobilisation d'une

technique existante, l'interprétation des codes comme facteurs, et leur manipulation effective suivant la technique adoptée.

La réassignation d'outils existants en vue de la résolution d'un problème donné relève de ce que Gilbert Simondon définit non pas comme l'inventivité, comme nous l'avons suggéré plus haut, mais comme la créativité, c'est-à-dire la capacité à ne pas limiter les propriétés des objets techniques à leur usage général, et à en dégager de nouvelles propriétés leur permettant de jouer un rôle instrumental et médiateur dans la situation, qui est un élément clé de la définition. Ce rôle consiste à servir de pont opérationnel entre la situation présente et l'image mentale du résultat visé formée par la personne créative par anticipation. Comme le précise Simondon dans *L'invention dans les techniques* (2005), la créativité se distingue ainsi de l'intelligence par son rapport à une situation donnée :

> L'intelligence peut être présentée comme une aptitude générale à résoudre les problèmes, tandis que l'intelligence créative se manifeste non pas en tous les problèmes, mais dans ceux où la solution est rendue possible par la sélection d'un cas particulier ou par celle d'une médiation unique particulièrement rapide et efficace, parmi plusieurs solutions possibles.

La créativité se distingue également de l'invention, à laquelle il donne un sens particulier, et dont il propose une théorie dans laquelle l'invention procède d'une méthode raisonnée constituée de plusieurs phases distinctes, sur lesquelles nous allons revenir. Une solution créative est en effet une solution nouvelle, mais c'est également une solution unique : ce qui a pu fonctionner dans ce cas particulier, dans cette situation donnée, ne pourra pas être généralisé ou répété en l'état. La créativité développe sa propre logique, sa propre rationalité, bien qu'elle puisse mobiliser d'autres logiques en les instrumentalisant. Le récit permet de rendre compte de cette rationalité singulière et d'en restituer le sens. Telle que nous l'avons décrite, la configuration en récit correspond à la fois à une forme de restitution, la synthèse temporelle de l'hétérogène par le récit, et à une procédure de constitution, la créativité fondée sur la réassignation d'objets familiers.

La deuxième configuration des projets de traitement de données numériques est celle du **système**. Elle consiste à produire une forme computationnelle non linéaire, typiquement un logiciel applicatif, qui résulte d'un processus de conception et d'ingénierie logicielle, et dont le livrable a pour vocation de fonctionner de manière autonome. Il est composé d'instructions, d'éléments programmés, de données et d'éléments d'interfaçage avec son environnement technique (la machine, le système d'exploitation, les autres programmes, etc.) et humain (le développeur, l'utilisateur). Contrairement au récit, ce livrable n'est pas figé, d'une part parce qu'il permet des interactions utilisateur, et d'autre part parce qu'il est développé de manière itérative, et peut évoluer dans le temps. Ce caractère interactif délinéarise la manipulation de données et définit un espace de séquences d'actions possibles. En effet, le logiciel rend possible le traitement de données ; il est le transcendantal empirique qui détermine quelles opérations sur les données sont accessibles à l'utilisateur. Là où la configuration en récit relevait d'une situation singulière, la configuration en système prévoit un modèle d'usage

virtuellement compatible avec une infinité de situations, de moments singuliers et de gestes répétables. Ce modèle d'usage existe sur le mode de la recommandation : par le cadre qu'il définit, il suggère des actions possibles, mais ne les effectue pas, car c'est toujours l'utilisateur qui reste l'agent des décisions d'interaction avec lui.

Sous ces termes conceptuels et généraux, la configuration en système décrit en fait une multitude d'exemples très familiers. Les algorithmes qui émerveillent ou inquiètent tant les médias et leur public, ainsi qu'une partie de la communauté de chercheurs en sciences de la culture, relèvent tous ou presque de cette configuration en système. Le flux d'actualité de Facebook, le moteur de recherche de Google, la recommandation de produit d'Amazon, ou de produits culturels de Netflix, le ciblage publicitaire, et tant d'autres « algorithmes », sont en réalité des systèmes qui intègrent certes des procédures algorithmiques de traitement de données, mais aussi et peut-être surtout des suggestions d'action, une capacité d'interaction, et une logique de prise en compte rétroactive de cette interaction. Ces systèmes existent sur le mode de la suggestion car c'est l'utilisateur qui décide quel contenu il va commenter, quelle page web il va consulter, quel produit il va choisir[42] : ce sont tous des systèmes de recommandation.

Il nous faut souligner ici que considérer tous les systèmes fondés sur le traitement de données massives comme des systèmes de recommandation peut paraître provocateur au regard du nombre et de la diversité de ces systèmes. Ainsi Karen Yeung (2017) montre également que bon nombre de systèmes sont en effet subsumés sous cette définition, mais elle les distingue d'une seconde catégorie, les systèmes automatiques, qui ne produiraient pas des recommandations, mais des décisions. Néanmoins, lorsqu'on examine attentivement son analyse de ces systèmes, on observe que ses exemples sont rares et similaires : radars automatiques, régulation des feux de signalisation, détection d'une fraude sur une carte bleue, détection d'une tentative d'intrusion dans un système d'information, etc. Dans tous les cas, il s'agit de systèmes, ou plus précisément de sous-systèmes, de régulation et de contrôle, qui ne sont pas construits en vue d'un utilisateur, mais seulement d'un administrateur. Certains de ces sous-systèmes de contrôle existent d'ailleurs sous le mode de la suggestion et non de la décision automatique : le système estime qu'il peut y avoir fraude, mais ne déclenche pas lui-même une sanction. De plus, qu'ils prennent ou non une décision automatisée, ces sous-systèmes ne remplissent pas la fonction première du système, mais s'assurent que cette fonction est assurée comme prévu. Ainsi, la fonction d'un système de paiement par carte bancaire n'est pas de détecter la fraude, objectif secondaire, mais de permettre le paiement d'un utilisateur. Ces sous-systèmes de contrôle et de régulation sont secondaires dans des systèmes plus larges qui produisent eux de la recommandation en vue d'un utilisateur.

[42] Soulignons ici qu'en ce qui concerne les moteurs de recherche, il y a en réalité deux décisions : la première relative aux critères de recherches (termes, restriction de date, type de contenu, etc.) et la seconde relative à la page web visitée par l'utilisateur. C'est sur ce second choix que nous nous concentrons pour l'examen des systèmes de recommandation.

Pour produire ces suggestions, tous ces systèmes de recommandation fabriquent du mesurable, c'est-à-dire qu'ils établissent des distances (ou ressemblances) entre éléments à recommander, entre utilisateurs, et entre les éléments et les utilisateurs. Toute suggestion est le résultat d'une fonction mathématique qui minimise une de ces distances ou une combinaison de celles-ci. Pour calculer effectivement cette distance minimale, le *data scientist* qui programme la recommandation a à sa disposition l'ensemble des techniques computationnelles issues des lignées que nous avons présentées au chapitre V, et la capacité à réassigner ces techniques à son projet épistémique indifféremment de leur lignée d'origine.

La recommandation comme mode de présentation des connaissances soulève également des questions relatives à la nature de l'action dans ce contexte précis. Ainsi, on pourrait remarquer que l'utilisateur n'a vraiment pas le choix, que ce choix est prédéterminé et orienté, que l'utilisateur est enfermé dans un nombre limité d'options. C'est bien effet le sens des notions de suggestion, ou de recommandation : si l'utilisateur reste l'agent du choix, ce n'est pas lui qui formule les différentes options parmi lesquelles il peut choisir. Or, l'établissement ou la hiérarchisation de ces différentes options peut avoir de lourdes conséquences morales relatives non seulement à la liberté de l'utilisateur, mais aussi à l'influence de ces choix programmés sur ses préférences et son agir.

Cependant, conceptuellement et techniquement, l'utilisateur se voit d'une part présenter plusieurs options, et peut d'autre part élire l'une (ou plusieurs) d'entre elles. Il possède au sein du système la capacité indivisible, atomique, à décider entre plusieurs options. Cette distinction, si ténue soit-elle quand le choix est limité, est ce qui sépare un système de recommandation, d'un système de décision, où il n'y a pas d'utilisateur humain agent de la décision. Ainsi, les systèmes de recommandation sont autonomes dans la mesure où ils n'ont pas besoin d'un développeur pour continuer à fonctionner[43] ; en revanche, ils ne sont pas indépendants dans la mesure où ils appartiennent à un système de relations et d'interaction plus large, de nature différente, qui intègre le logiciel mais aussi l'utilisateur sans lequel ce système plus large est incomplet et inactif.

Enfin, soulignons que, du fait que ces logiciels intègrent des logiques d'apprentissage automatique, l'(inter)action utilisateur est le déclencheur d'une réaction du système (un contenu est publié, un produit est acheté, une page s'affiche, etc.) mais constitue aussi elle-même une donnée supplémentaire, intégrée par rétroaction aux procédures de traitement de données qui prédéterminent les recommandations futures du système. Trois éléments constituent ainsi l'activité de traitement de données du point de vue de l'utilisateur : les propositions du système, la décision de l'utilisateur, et la rétroaction pour l'apprentissage. Toutefois, ces trois éléments restent associés à l'utilisateur final ; ils sont la forme de restitution du résultat du traitement de données. Aussi, de même que la configuration en récit pouvait être approchée par sa procédure de constitution comme par sa forme de restitution,

[43] Ils ont, en revanche, souvent besoin d'administrateurs système, car un logiciel informatique est très sujet aux pannes.

la configuration en système peut également être abordée *via* ses résultats, mais aussi *via* la généalogie du système lui-même.

2.2 La genèse des systèmes

La généalogie d'un système de recommandation nous conduit dans celle de l'industrie logicielle en général. On appellera ainsi moteur de recommandation la fonctionnalité qui se charge de calculer des choix pour l'utilisateur ; la prise en compte rétroactive de la décision de l'utilisateur et permettant d'améliorer ses résultats est une sous-fonctionnalité du moteur. Ce moteur s'insère dans une écologie, c'est-à-dire un environnement technique composé d'éléments matériels, logiciels, d'autres moteurs, d'interfaces, de tâches programmées, de bases de données, d'un réseau, de contraintes temporelles et computationnelles, etc. L'élément « moteur de recommandation » peut se conceptualiser à un niveau théorique, comme une fonction qui minimise des distances entre éléments. Il peut se décomposer en tâches élémentaires qui seront semblables d'un système à l'autre, utiliser des techniques semblables comme les algorithmes de filtrage collaboratif *(collaborative filtering)*. À l'inverse, son écologie et son implémentation relèvent d'un assemblage singulier dont les modalités peuvent difficilement être généralisées. Dans chaque implémentation (c'est-à-dire chaque système de recommandation effectivement existant), le schème conceptuel du moteur de recommandation fait l'objet d'un processus d'individuation et de concrétisation au sens du *Mode d'existence des objets techniques* (Simondon 1958). Elle est une réponse *sur mesure* au problème de créer du répétable.

La mobilisation de Simondon est ici porteuse d'intelligibilité car l'implémentation d'un système de recommandation est exactement un objet technique « dont il y a genèse spécifique procédant de l'abstrait au concret ». Le moteur de recommandation au niveau conceptuel joue le rôle de la représentation scientifique dont l'implémentation est le passage de l'abstrait (une fonction de minimisation) au concret (un objet technique inséré dans son milieu). L'obsession de la figure de l'informaticien pour l'optimisation, qui nous avions mise en évidence au chapitre V à travers la notion de pensée computationnelle, est le moteur psychologique du processus de perfectionnement et de concrétisation qui constitue la genèse de l'objet technique : il permet l'adaptation de l'objet à lui-même (milieu interne) et à son environnement technique (milieu externe). Ce perfectionnement a lieu au niveau de son milieu interne par l'amélioration de la fonction de minimisation d'une distance, et au niveau de son milieu externe, par réduction des frictions (lenteurs et erreurs) dans les interactions avec son environnement technique. Plus le code est simple, sans détour, efficace, plus il est considéré comme techniquement perfectionné.

La mobilisation de Simondon sur des objets numériques ne peut en revanche pas se faire telle quelle : lorsqu'il écrit *Du mode d'existence des objets techniques* (*MEOT*), Simondon est contemporain du développement de la cybernétique et des machines à information qui préfigurent l'informatique contemporaine, mais ses exemples proviennent surtout de la conception industrielle et notamment de

la mécanique et de l'électronique. Les illustrations de la première partie de *MEOT* (la diode, la turbine, le moteur, l'automobile, etc.) sont résolument antérieurs à l'ère numérique ; ce sont des objets industriels, manufacturés, dont le substrat est matériel. Il s'intéresse davantage à leur invention qu'à leur manufacture. Leur production relève du génie mécanique, là où les moteurs de recommandation relèvent du génie logiciel.

La spécificité du génie logiciel par rapport au génie classique, qu'il soit civil, mécanique ou chimique par exemple, est que le produit peut se transformer en continu du fait de son substrat numérique : les phases de conception et de fabrication ne sont presque pas séparées. Cette affirmation est moins vraie des systèmes embarqués, associés à un substrat physique spécifique, où les développeurs n'ont plus accès au logiciel après l'avoir livré (ou difficilement, par un rappel produit massif par exemple) mais elle décrit en revanche particulièrement bien les applications web. Au contraire du pont, de l'automobile, du transistor, qu'il faut fabriquer à nouveau si l'on souhaite qu'il soit différent, le logiciel est toujours (presque) entièrement modifiable. On notera d'ailleurs ici qu'il y a une différence de taille entre le processus de construction d'un pont (objet unique) et celui d'un transistor, productible en masse d'un point de vue synchronique, et évolutif d'une génération à l'autre d'un point de vue diachronique. Entre l'émergence du génie civil pour les grands travaux publics, à partir du génie militaire, vers la moitié du XIXe siècle, et le développement des industries chimiques et électroniques au XXe siècle (Fauré-Fremiet 2017; Rodrigues 2017), il y a déjà deux figures de l'ingénieur que l'on peut distinguer, mais qui relèvent globalement du même ethos, que nous détaillerons plus bas. Dans le génie logiciel, il n'est pas nécessaire de modifier physiquement des chaînes de production pour permettre la prise en compte d'une amélioration, contrairement au contexte classique où cette prise en compte s'inscrit dans une économie technique et financière conséquente. La reproductibilité de l'objet numérique est toujours déjà virtuellement infinie et instantanée (bien que son coût ne soit pas complètement nul). De ce fait, sa genèse est différente de celle des objets physiques ; en particulier, on ne peut pas dissocier, comme le fait Simondon, les phases d'invention et de fabrication.

En effet, il n'y a plus une phase de conception puis une phase de production, mais des phases de conception et d'implémentation généralement fortement imbriquées, où les choix de conception sont faits avant et pendant la programmation informatique, et une phase de mise en production, quasi-instantanée, qui consiste à publier le code. Soulignons ici qu'un informaticien professionnel n'est pas seulement quelqu'un qui écrit du code : indépendamment des tâches liées à la gestion et la communication au sein d'une entreprise (ou d'une autre organisation), son travail se vit également comme un exercice de réflexion et de recherche de solution. Par conception on entend donc ici ce travail de réflexion qui consiste également à réfléchir à la structure du programme, son architecture, la façon dont il ordonne ses opérations successives, et son intégration dans l'environnement technique ; l'implémentation correspond alors à l'écriture effective du code, qui soulève souvent de nouvelles questions de conception au fil du curseur, d'où une imbrication forte de ces deux phases

séparées un peu artificiellement ici. Le programme s'apparente moins, conceptuellement, au produit manufacturé du génie industriel, concrétisé et reproductible, qu'à la chaîne de production, c'est-à-dire de l'élément qui doit pouvoir produire de manière fiable et répétable le même résultat indéfiniment, suivant le schème défini par l'objet technique. Néanmoins, le parallèle avec les concepts simondoniens touche ici sa limite, car la concrétisation de l'objet technique se fait autrement que par invention successive ; ils restent pertinents, cependant, pour penser l'individuation des objets techniques (l'algorithme, le programme, le moteur, le système, etc.) par rapport à leur milieu.

La genèse d'un système de recommandation comporte grossièrement deux phases, dont une première qui, si elle relève bien du milieu informatique, appartient en revanche rarement au génie logiciel. En effet, le secteur informatique est caractérisé par un va-et-vient permanent entre l'artisanat et l'industrie (alors que les concepts simondoniens sont formulés en ayant à l'esprit l'ère industrielle et deviennent, ici encore, difficiles à appliquer). Le moteur de recommandation développé par une organisation commence généralement par exister avec un statut expérimental ; sa réalisation, une fois dépassée la première forme, sommaire, où l'on se rend compte qu'on ne peut pas « bricoler » une solution qui fonctionnera efficacement dans tous les cas, est d'abord la responsabilité d'un type d'équipe qui peut avoir des noms variés, mais relève d'une fonction « recherche et développement ». Ce type d'équipe produit des solutions plus complexes qu'un bricolage sommaire, mais sous une forme plus artisanale qu'industrialisée, qui n'est pas à même de passer en production.

En effet, lorsque nous avons évoqué jusqu'ici la figure de l'informaticien, nous avons passé sous silence la complexité des métiers et des fonctions directement en lien avec l'informatique. Rappelons ici que l'informatique est un secteur d'activité économique à part entière, qui emploie des millions de personnes à travers le monde, avec une diversité de profils qui soulève par ailleurs de réelles difficultés aux recruteurs. Ces profils ont des formations de niveau et de longueur variés (autodidacte, technicien, ingénieur, docteur, etc.), des spécialités différentes (développement, intégration, administration, réseau, etc.), et des fonctions variables au sein de l'entreprise (développement logiciel, support, maintenance, conseil gestion de projet technique, etc.). C'est aussi un domaine de recherche, qui innove fréquemment, ce qui a pour conséquence de transformer en permanence les métiers qui relèvent de cette compétence.

Le titre d'ingénieur crée une confusion supplémentaire dans cet environnement. En France notamment, le terme est chargé d'une histoire particulière, retracée notamment par Theodore Porter (1995). Les grandes écoles d'ingénieurs, historiquement créées pour former les membres des grands Corps de l'État, ont fondé cette figure par opposition à celle du scientifique. Initialement très théorique et centrée sur les mathématiques, la formation d'ingénieur (civil) prend le sens que nous lui donnons aujourd'hui au début du XIXe siècle, notamment sous l'impulsion de l'École Polytechnique et de l'École des Ponts et Chaussées. En 1819, le Conseil de l'École Polytechnique déclare ainsi :

> Quand on considère le développement que prend tous les jours l'industrie en France, et qu'on envisage les rapports nécessaires de cette industrie avec la forme de gouvernement établie par la Charte, on doit sentir que l'exécution des travaux publics tendra, dans un très-grand nombre de cas, à passer dans le système de concession et d'entreprise. Il faut donc désormais que nos ingénieurs sachent régler et diriger ce mouvement. Il faut qu'il sachent évaluer l'utilité ou l'inconvénient particulier ou général de telle ou telle entreprise ; il faut par conséquent qu'ils aient des idées justes et précises sur les élémens [sic] de toutes ces spéculations, c'est-à-dire sur les intérêts généraux de l'industrie et de l'agriculture, sur la nature et l'influence des monnaies, sur les emprunts, les assurances, les fonds d'association, d'amortissement, en un mot sur tout ce qui peut servir à apprécier les bénéfices et les charges probables de toutes les entreprises: tel est l'ensemble des objets qui viennent d'être ajoutés au programme. (Fourcy, 1828, cité par Porter, 1995)

Cette citation, volontairement reproduite ici de manière extensive, est le reflet d'un tournant dans la formation d'ingénieur qui n'est pas sans faire écho à celui que les grandes écoles traversent aujourd'hui (bien qu'en réalité l'École Polytechnique ait largement conservé sa prédilection pour les formations abstraites et notamment les mathématiques[44]). En 1819, l'accent est mis sur l'importance des capacités d'évaluation et de jugement, des connaissances générales du contexte (notamment économique) d'un projet, bref, différents éléments prenant en compte ce que nous avons appelé l'écologie d'un projet. Comme le souligne Maria de Lurdes Rodrigues (2017) dans un article sur la profession d'ingénieur au Portugal :

> les ingénieurs sont non seulement des hommes de la technique, mais aussi des hommes de la gestion et de l'organisation – autrement dit, les dimensions relationnelles, administratives et de direction, dans l'activité des ingénieurs, sont aussi importantes que la dimension technique. Ce sont les hommes des organisations (entreprises, administrations publiques, banque, services) préparés pour exercer des fonctions d'encadrement hiérarchique ou en équipe et entraînés à trouver des solutions, compte tenu des contraintes techniques et économiques, à des problèmes qui ne sont pas purement techniques.

Par ailleurs, la figure de l'ingénieur est associée à un ensemble de valeurs morales et à une forme de responsabilité devant la nature et la société, lisibles par exemple dans l'engagement prononcé par les ingénieurs canadiens depuis le début du XXe siècle[45] :

> As an engineer, I, (full name), pledge to practice Integrity and Fair Dealing, Tolerance, and Respect, and to uphold devotion to the standards and dignity of my profession, conscious always that my skill carries with it the obligation to serve humanity by making best use of the Earth's precious wealth.
>
> As an engineer, I shall participate in none but honest enterprises. When needed, my skill and knowledge shall be given without reservation for the public good. In the performance of duty, and in fidelity to my profession, I shall give the utmost.

[44] Comme le souligne encore Theodore Porter (1995), "If Polytechnique was a school of engineering it was no institution of nuts and bolts, or gravel and paving stones. Engineering at the Ecole Polytechnique was as abstract and mathematical as the study of roads and bridges or artillery could possibly be, and possibly even more so."
[45] http://www.ironring.ca/contexte.php (consulté le 22 mars 2017)

Les ingénieurs canadiens appartiennent à un ordre, à l'instar des médecins ou des avocats. Leur profession renvoie à des compétences spécifiques, et notamment la capacité à prendre en compte le contexte et les contraintes techniques et non techniques d'un projet d'ingénierie, mais aussi à des obligations morales dont le non-respect peut leur valoir une interdiction d'exercer. Cette responsabilité morale procède à la fois du sentiment d'être investi d'un rôle social et politique (Diogo 2017), et d'une responsabilité davantage liés aux risques des chantiers et des constructions, notamment dans des domaines comme le nucléaire.

À l'inverse, les grandes écoles françaises dispensent aujourd'hui des formations où l'informatique occupe une large place, suivant quoi on peut dire qu'elles forment des informaticiens. Néanmoins, bien que les étudiants obtiennent un titre d'ingénieur à l'issu de leur formation, le fait qu'elles forment des *ingénieurs en informatique* est soumis à discussion[46]. La formation en informatique des grandes écoles renvoie essentiellement à un ensemble de compétences théoriques et techniques liées à la conception et l'implémentation de programmes informatiques ; les écoles d'ingénieurs forment à ce titre de bons programmeurs. La capacité à penser l'architecture d'un système et son écologie sont vues comme des compétences plus tardives, qui viennent avec l'expérience ; les méthodes de conception et de développement logiciel sont elles aussi superficiellement enseignées au mieux. Nous voulons donc soutenir que s'il existe bien des développeurs, des techniciens, des intégrateurs, des consultants en informatique, etc., les *ingénieurs en informatique*, et plus spécifiquement, des *ingénieurs en développement logiciel* (en anglais *software engineer*), sont moins nombreux, et qu'il s'agit davantage, au-delà de la sanction du diplôme, d'un devenir auquel on parvient avec l'expérience.

En particulier, les équipes de type « recherche et développement » produisent des solutions qui, comme nous le disions, ne sont pas à même de passer en production telles quelles, et ne constituent pas un travail d'ingénierie logicielle au sens où nous l'entendons, bien qu'ils s'apparentent bien à la figure de l'ingénieur bricoleur et créatif. Il s'agit toutefois d'équipes qui rassemblent des profils avec d'importantes compétences techniques, des niveaux de diplôme et de prestige de formation supérieurs à la moyenne des informaticiens. C'est notamment dans ce type d'équipe que l'on rencontre la plus forte proportion de docteurs, non seulement en informatique, mais aussi en physique, chimie, économie, et plus largement issus de domaines scientifiques fortement mathématisés. Ces profils ont en effet en commun de mieux maîtriser les techniques d'algorithmique et d'apprentissage automatique à l'œuvre dans les sciences des données, ainsi que les concepts mathématiques sous-

[46] Y compris parmi les informaticiens eux-mêmes, voir par exemple :
- "You are not a software engineer" (mai 2011), http://www.chrisaitchison.com/2011/05/03/you-are-not-a-software-engineer/ ;
- "You think you're an engineer but you're not" (avril 2013), http://nic.ferrier.me.uk/blog/2013_04/you-are-not-an-engineer ;
- "Programmers: Stop Calling Yourselves Engineers" (novembre 2015), https://www.theatlantic.com/technology/archive/2015/11/programmers-should-not-call-themselves-engineers/414271/ ;

(consultés le 22 mai 2017).

jacents. Idéalement, ils détiennent une multitude de compétences (programmation, communication, graphisme, connaissances métier) dont la liste ne cesse de s'allonger (voir **Figure 14**) jusqu'à l'agacement ou l'amusement des intéressés. En pratique, les profils qui présentent toutes ces compétences sont très rares, ou inaccessibles aux entreprises du fait de leurs exigences salariales ; les *data scientists* ont le plus souvent de bonnes connaissances en mathématiques appliquées, et des compétences plus ou moins sommaires en programmation. De plus, ils maîtrisent généralement mieux des langages de programmation plutôt conçus pour le calcul mathématique ou l'analyse statistique, comme MATLAB et surtout R, que des langages privilégiés pour les systèmes d'information et les architectures complexes tels que le C++, le Java ou le C#.

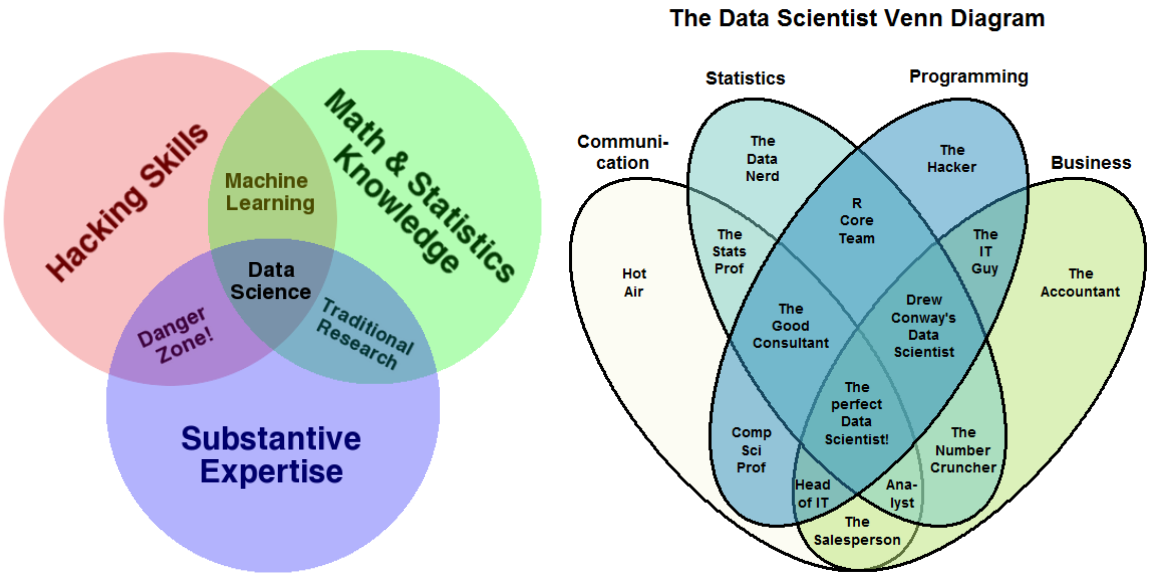

***Figure 14**. Deux visions de la combinaison de compétences du data scientist idéal, la première par le consultant Drew Conway[47], et la seconde par Dr Stephan Kolassa[48], « Data Science Expert ».*

De ce fait, si la contribution du *data scientist* est nécessaire à la genèse des systèmes de recommandation, elle est d'une nature spécifique, centrée sur le schème fonctionnel interne du moteur de recommandation plutôt que sur son intégration dans le système. En des termes simondoniens, nous dirons qu'ils sont davantage acteurs de l'autocorrélation de l'objet technique que de sa médiation : ils agissent sur son perfectionnement pour ainsi dire *in abstracto*, indépendamment de son milieu technique. Cette observation s'explique pour plusieurs raisons, dont la première est celle que nous avons développée ci-dessus concernant la figure du *data scientist* : la prise en compte de l'écologie de l'objet est rarement enseignée comme une nécessité dans les formations d'ingénieurs, et tout aussi rarement abordée dans les projets de recherche menés par les doctorants en informatique.

[47] http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram (consulté le 23 mai 2017)
[48] https://datascience.stackexchange.com/questions/2403/data-science-without-knowledge-of-a-specific-topic-is-it-worth-pursuing-as-a-ca (consulté le 23 mai 2017)

Elle procède également de la situation dans laquelle ils travaillent. En effet, de par leur fonction dans l'entreprise, ils n'interviennent pas sur le système de recommandation visible par les utilisateurs, mais sur des jeux de données extraits de celui-ci, qui servent à l'expérimentation. Il s'agit ainsi d'un travail de recherche « in vitro » où les *data scientists* peuvent se concentrer sur l'autocorrélation du moteur de recommandation et mobiliser les multiples familles de techniques algorithmiques présentées au chapitre V au service de la fonction de minimisation de distance entre éléments. Cette configuration comporte un risque associé à l'extraction du jeu de données, qui peut s'avérer de plus en plus en décalage avec les nouvelles données utilisées en production.

Par ailleurs, elle n'est pas sans rappeler, en réalité, la configuration en récit, dans la mesure où les *data scientists* travaillent sur une situation spécifique, déterminée par le jeu de données, dont ils rendent compte à leurs collègues avec le code produit, mais aussi avec une forme de rapport oral ou écrit qui retrace leur processus de travail. Pour faire aboutir ce travail, ils commencent généralement par adopter une approche exploratoire, qui mobilise la visualisation de données, à la fois pour se donner un aperçu des données, et contrôler les résultats de leurs calculs. De plus, leur travail relève tantôt de la créativité (dont on a vu qu'elle était caractéristique de la configuration en récit) et tantôt de l'invention : ainsi, comme l'analyste de la configuration en récit, ils mobilisent des techniques computationnelles préexistantes et les combinent pour former une solution qui s'apparente plus à un bricolage qu'à une invention. Les techniques d'apprentissage profond en particulier exigent de nombreux ajustements pour donner des résultats convaincants. Ces ajustements relèvent du tâtonnement et du bricolage plus que du perfectionnement mathématique ou computationnel ; il n'en demeure pas moins vrai que la qualité des résultats est très sensible à ces ajustements.

À l'inverse, si les techniques qui permettent de constituer un moteur de recommandation constituent en effet une forme de boîte à outils qui peut être réutilisée dans des situations variées, le *data scientist* doit cependant donner forme à une solution qui correspond à un problème plus spécifique le problème de la recommandation en général : ce n'est pas la même chose que de proposer des contenus, des objets physiques ou des produits culturels. Chaque entreprise se caractérise également par une certaine approche du problème : en termes de recommandation musicale par exemple, des organisations comme Deezer, Spotify et last.fm ont des stratégies différentes. À ce titre, le travail du *data scientist* ne consiste pas toujours à « bricoler » une solution, mais à en inventer une, à formuler une proposition singulière et nouvelle par rapport à un problème donné.

Cette proposition circule ensuite en dehors des équipes de recherche et développement, bien qu'elle puisse être amenée à y revenir, et à faire des allers-retours. Pour passer en production et arriver aux mains (ou aux yeux) des utilisateurs, elle va faire l'objet de traitements qui relèvent de l'ingénierie logicielle à proprement parler, c'est-à-dire un travail de conception et d'implémentation technique prenant en compte l'écologie du projet, ou ce que Simondon appellerait le milieu associé. Comme toute ingénierie, elle s'organise autour de schémas, de techniques et de méthodes de développement

qui structurent son activité. Les méthodes dites « agiles » s'inscrivent ainsi en rupture des méthodes de production industrielle tout en puisant dans leurs innovations. Ces méthodes n'englobent pas seulement le développement technique des produits logiciels, mais tendent à structurer l'intégralité des entreprises technologiques dont l'activité est centrée sur le développement logiciel. Ainsi la méthode « lean startup » présentée par Eric Ries dans son best-seller *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses* (2011) n'est pas tant une méthode de développement logiciel qu'une méthode de gestion et d'organisation à destination des entreprises qui développent des produits basés sur des technologies informatiques, par ailleurs inspirée de l'organisation de la production automobile japonaise, et notamment de Toyota. Les promoteurs de ces méthodes tendent d'ailleurs systématiquement à étendre leur domaine, d'un outil méthodologique de gestion d'équipe à des principes de structuration et de fonctionnement de toute l'entreprise, voire à une philosophie de vie mêlée de principes moraux.

Plus concrètement, le noyau de toutes ces méthodes est de fonctionner sur des cycles rapides et successifs de développement et d'évaluation. Dans la méthode SCRUM, le travail d'une équipe est décomposé en *sprints*, des périodes qui peuvent aller de quelques heures à quelques semaines, au terme desquelles le *scrum master* évalue le travail produit et planifie d'autres sprints, sur la base de décisions souvent prises collégialement. L'orientation générale du logiciel est donc réévaluée rapidement, sans que tout son contenu n'ait besoin d'être défini préalablement. Ces approches, qui permettent d'obtenir des *livrables* et d'ajuster les objectifs sans attendre la fin du développement du produit complet, ont été mises au point notamment en réaction aux cycles de développement classiques comme le cycle en V devenu la norme des grands projets informatiques dans les années 1980, où l'évaluation n'était possible qu'au terme du projet, et lui-même hérité du fonctionnement en cascades issu du génie civil.

À ces méthodes d'organisation et de gestion des équipes s'ajoutent des méthodes de développement informatique à proprement parler, comme par exemple le développement piloté par les tests *(test-driven development)* qui consiste à écrire les tests unitaires d'une fonctionnalité avant le code correspondant. Un test unitaire est lui aussi un programme, qui permet de contrôler un aspect précis de la fonctionnalité en vérifiant automatiquement, par un test binaire, qu'elle produit bien le résultat attendu ; une même fonctionnalité peut comporter plusieurs dizaines de tests unitaires, dont le code correspondant peut être plus long que celui qui implémente la fonctionnalité elle-même. Ces tests unitaires sont exécutés par des outils d'intégration continue qui automatisent leur exécution ainsi que celle d'autres tâches comme la compilation du code, les tests de performance, de manière à contrôler régulièrement la stabilité des fonctionnalités du logiciel. Il existe également des méthodes de structuration du logiciel au niveau micro et macro ; les patrons d'architecture et de conception *(design pattern)* définissent des bonnes pratiques d'agencement des modules, des fonctions, des classes, etc. Enfin, des outils et méthodes régissent non pas l'écriture du code mais sa circulation : par exemple,

les outils de gestion de versions comme Git permettent de contribuer à plusieurs sur un projet et de maîtriser son évolution dans le temps.

Tous ces éléments de structuration de la production logicielle vont dans le sens d'une ingénierie développée, à la fois distincte et apparentée aux formes classiques de génie. Contrairement aux projets d'expérimentation ou de recherche individuels, le projet de créer et de faire évoluer un logiciel peut difficilement s'exécuter avec succès sans cette structuration méthodologique et instrumentale. La figure du *software engineer* dessine bien une figure d'ingénieur concepteur, généralement expérimenté, à laquelle s'oppose la figure de l'ingénieur bricoleur ou *hacker*.

Bien qu'il existe des exceptions, et que nous avons dessiné deux figures en opposition là où elles forment plutôt un spectre sur lequel s'inscrit tout développeur, la figure de l'ingénieur capable d'industrialiser et de rendre robuste un système se rattache tout particulièrement à la phase de développement logiciel qui suit la phase de conception du moteur de recommandation, dans la configuration en système. En d'autres termes, lorsqu'un analyste mobilise un langage de programmation et ses bibliothèques plutôt que des outils à interface graphique, il le fait sur le mode du bricoleur, qui écrit un code jetable (un *script)*, prévu pour n'être utilisé qu'une fois. C'est typiquement le mode du consultant en sciences des données, qui va proposer une démonstration qui fonctionnera jusqu'à la conclusion du projet, et rarement au-delà. Il en va de même du *data scientist* dans les départements R&D qui travaille, souvent de manière autonome, sur le développement d'un moteur de recommandation apprenant. Dans les deux cas, il travaille sur une situation définie par les données et la problématique. De ce fait, quand bien même il adopterait les méthodes et outils du génie logiciel, il ne peut, par définition, pas intervenir sur la prise en compte de l'environnement technique qui devient nécessaire lorsque les traitements de données ne se font plus *in vitro* mais *in vivo*, ou autrement dit, en production.

Dans le monde du logiciel web, l'environnement de production est celui qui est accessible aux utilisateurs. C'est la « vraie » version du logiciel, la plus définitive, la plus à jour, celle qui ne doit pas tomber en panne. Cette définition est valable des applications et services web, où l'utilisateur se connecte au logiciel installé sur cet environnement, mais aussi des versions téléchargeables. Cet environnement de production s'oppose aux autres environnements à usage interne ou semi-interne, tels que les environnements de test, de recette, d'intégration, de validation. La phase d'intégration du moteur de recommandation dans son système vise à la fois à anticiper les interactions entre le moteur et les autres composants du système, et à le préparer à son passage en production.

L'une des préoccupations majeures de l'ingénierie logicielle vis-à-vis des contributions des sciences des données est celle du passage à l'échelle ou scalabilité. Le programme qui fonctionne dans une situation connue, spécifique et maîtrisée, est dans une configuration de fragilité s'il est déployé pour des milliers d'utilisateurs, pour une durée indéterminée, sur un environnement différent. Une bonne

partie des technologies commercialisées avec le libellé « *big data* » servent ainsi à répondre à la problématique de réaliser une opération computationnelle courante, comme un tri ou une agrégation, mais sur plusieurs serveurs, et dont les particularités sont inconnues au moment du développement. Sans forcément basculer sur une architecture distribuée, il peut être nécessaire par exemple de réécrire le code du moteur de recommandation dans un autre langage, plus performant du point de vue des temps d'exécution des programmes. Par ailleurs, un calcul pourra fonctionner sur l'échantillon de données exploité par le data scientist pour mettre au point son moteur, mais pas sur les données de production. Il est susceptible de proposer de bonnes suggestions, mais après un délai inacceptable pour un utilisateur. La plupart des systèmes de recommandation ne sont acceptables que si leurs suggestions sont virtuellement instantanées : on n'imagine pas devoir attendre plusieurs heures ni même plusieurs minutes avant de pouvoir consulter son fil Facebook ou les films suggérés par Netflix. Le temps du logiciel fait donc lui aussi l'objet d'une ingénierie : certains calculs seront simplifiés pour pouvoir être exécutés « à la volée », lorsque l'utilisateur clique sur un bouton ou entre une requête ; d'autres seront décorrélés du temps de l'utilisation, exécutés en général la nuit, parfois sur d'autres serveurs, par l'intermédiaire de tâches programmées qui appartiennent à leur tour à des systèmes de planification et de répartition de la charge de calcul. Les services de recommandation musicale proposent ainsi plusieurs types de suggestions : des suggestions quotidiennes, ou hebdomadaires, dont le contenu est calculé dans un temps différent du temps de l'écoute, et des suggestions « à la volée » qui peuvent par ailleurs également s'appuyer sur le résultat de calculs plus longs.

Pour devenir un *système* de recommandation à proprement parler, le *moteur* passe donc par un ensemble d'étapes, suivant une méthode raisonnée qui n'est pas sans rappeler la conceptualisation de l'invention chez Simondon. La phase de prototypage par un *data scientist* se rapproche du processus d'analyse de données dans la configuration en récit, tandis que l'arrivée du prototype aux mains d'ingénieurs en développement logiciel le prépare à une répétabilité dans une multitude de situations qui doivent être anticipées. Ces deux phases d'adaptation interne et externe de l'objet technique sont nécessaires à son fonctionnement effectif. Cette répétabilité est mise en œuvre par le passage en production, où le système devient effectivement accessible aux utilisateurs. Pour eux, le système est un instrument avec lequel ils peuvent interagir, notamment en choisissant parmi des suggestions proposées par le système.

Conclusion

De l'algorithme de recommandation, issu d'une technique ou d'un agrégat de techniques présenté dans une revue scientifique, efficace sur une situation *in vitro* spécifique, à la fonctionnalité de recommandation intégrée dans son système, il y a une montée en complexité qui est le propre non pas des sciences des données, mais de l'ingénierie logicielle, au sein de laquelle le poids de l'algorithme est contrebalancé par la multiplicité des composantes techniques et humaines du système, et de leurs

interactions. Face à cette complexité, l'ingénierie logicielle permet de développer non pas l'infaillibilité des systèmes, c'est-à-dire leur capacité à empêcher l'accident, mais la robustesse, c'est-à-dire la capacité à fonctionner malgré l'accident, à l'incorporer, à l'ignorer. Elle fait du système une réponse non plus à la situation singulière définie par les données et la problématique, mais une catégorie de situations pouvant correspondre virtuellement à une infinité de situations réelles. Les techniques computationnelles des sciences des données y jouent un rôle nécessaire, car sans elle, la fonctionnalité n'existe pas, mais loin d'être suffisant dans l'écologie du projet constituée par l'industrialisation de la fonctionnalité dans le système.

La configuration en système se distingue ainsi de la configuration en récit par la forme de restitution qu'elle propose, comme on l'a vu, aussi bien que par sa méthode de constitution, qui se déploie bien au-delà du domaine d'action des sciences des données. Néanmoins, la genèse de ces systèmes passe presque toujours par une phase plus artisanale qui relève davantage de la configuration en récit, et hérite du rôle qu'elle donne à la créativité : on peut dire que la configuration en système intègre l'autre forme. En pratique, il y a des circulations continues entre ces deux configurations, ainsi qu'entre les valeurs épistémiques, les procédures et les méthodes qu'elles mobilisent. Tout comme l'opposition entre le bricoleur et l'ingénieur représentent les deux extrémités d'un spectre, qui sont par ailleurs loin de rendre compte de toute la complexité du métier de développeur, l'opposition entre les deux configurations fournit deux concepts mis en tensions qui permettent de caractériser des situations concrètes.

Ainsi, la façon dont le logiciel implémente la méthode d'analyse de données chez Proxem relève non pas de l'intégration d'une configuration dans l'autre, mais de la transition du mode récit au mode système. Dans cette transition, on passe d'une situation singulière à une multitude de situations dont on peut extraire, par commensuration, les éléments communs, qui pourront alors être industrialisés dans le logiciel. Par ailleurs, la place occupée par le moteur de recommandation prend des places distinctes d'un exemple à l'autre. Chez Proxem, plusieurs formes de recommandation sont à l'œuvre, à la fois pour identifier des observables lexicaux potentiellement pertinents (suggestions d'annotations), pour les assembler (suggestions de regroupements), et pour les parcourir (moteur de recherche, autocomplétion et résultats de recherche) ; aucun ne constitue pour autant le cœur du système. Dans le cas de Google, c'est l'inverse : le logiciel n'est rien d'autre qu'un système qui produit des propositions de contenus web, et autour desquels s'articulent des fonctions annexes.

Au-delà des différences que nous avons mises en évidence entre les différentes configurations que suivent les processus de traitement de *big data*, il demeure ainsi entre elles des schèmes similaires, relatifs à la fois à la mécanicité computationnelle des résultats produits et à l'espace interprétatif proposé par leur matérialisation. L'espace d'actions possibles dessiné par l'interface fait écho à la manipulabilité des visualisations de données en vue de leur perfectionnement. La mécanique de traitement de données est une mécanique qui a du jeu ; ce jeu n'est pas un défaut mais un moment

nécessaire qui laisse place à l'intervention humaine créatrice de sens. L'interface, la visualisation, rend possible et facilite cette intervention. Dans chaque cas, il y a quelque chose qui est donné mécaniquement par le calcul, mais qui doit encore être transformé pour faire sens et permettre l'action, qui doit faire l'objet de jugements et de manipulations qui relèvent alternativement des idiosyncrasies de l'utilisateur et de l'expertise de l'analyste. C'est ainsi seulement que prennent forme des connaissances infra-conceptuelles, que la visualisation acquiert sa signification, et que la recommandation devient action. Nous examinons à présent plus spécifiquement le rôle joué par la visualisation de données dans la constitution de ces connaissances.

Chapitre VIII.
L'art de la visualisation de données

Comme nous l'avons déjà souligné plus haut, la visualisation de données joue un rôle déterminant à la fois dans la compréhension des données, c'est-à-dire dans l'interprétation par l'analyste des résultats des calculs opérés sur les données, et dans la transmission de cette compréhension. De ce fait, la représentation graphique est à la fois une **source d'intelligibilité** pour l'analyste, et un **levier rhétorique** pour communiquer et convaincre un destinataire. La représentation graphique n'est pas le propre des données massives, et ce n'est en rien une invention récente. Comme nous l'avons amplement vu à travers son rôle dans l'analyse exploratoire de données qui émerge dès les années 1960, notamment avec John Tukey, elle lui est largement antérieure. Les sciences de la culture, et tout particulièrement la géographie, ont développé un art de la matérialisation visuelle largement antérieur au développement de l'informatique, et a fortiori de sa massification ; ses usages vont par ailleurs bien au-delà du contexte de la pratique scientifique. Néanmoins, la densité des données et la complexité des instruments accroissent le besoin d'une médiation visuelle pour appréhender les résultats des traitements. Ce besoin de médiation prend des formes spécifiques dans le contexte des *big data*, tout en remobilisant ou en redonnant sens à des formes graphiques plus ou moins tombées en désuétude.

1. La sémiologie graphique

Dans son principe, le service rendu peut être décrit simplement : la visualisation simplifie et synthétise ce qui ne peut être embrassé d'un seul regard, qu'il s'agisse de l'œil « physique » ou de « l'œil de l'esprit » et de son appréhension intellectuelle de l'information. Comme nous l'avons vu au chapitre V, elle est notamment chez John Tukey une médiation précieuse pour explorer les données elles-mêmes, rendue tangible lorsqu'on compare la lecture des données (à supposer qu'on puisse les représenter textuellement), et l'examen de leur représentation visuelle.

Ainsi, pour évaluer une relation entre deux variables par exemple, je peux tenter de parcourir les données elles-mêmes. Dans leur livre *Doing data science* (2014), Rachel Schutt et Cathy O'Neil donnent

l'exemple de données comportementales sur un réseau social quelconque, qui enregistre notamment le temps passé sur le site par un utilisateur et le nombre d'amis auxquels il est connecté. Si l'on regarde à l'œil nu les premières valeurs prises par ces variables **(voir Tableau 6)**, il est difficile de se faire une idée de leur relation, même si on peut avoir une intuition à ce sujet.

7	276
3	43
4	82
6	136
10	417
9	269

***Tableau 6**. Premières lignes d'un jeu de données comportementales enregistrées sur un réseau social, représentant les variables « temps passé sur le site » et « nombre d'amis » d'un utilisateur. Le rapport entre les variables n'est pas aisément lisible.*

On notera ici que le tableau est lui aussi une représentation graphique, qui matérialise chez Goody (1979) une catégorie conceptuelle, une structure de la pensée distincte de la lecture linéaire. Néanmoins, si cette représentation joue un rôle de structuration, par laquelle chaque élément tire son sens de son contenu mais aussi de sa position, elle n'effectue pas d'agrégation du contenu. À l'inverse, si l'on représente graphiquement non plus les premières lignes mais l'ensemble des données avec une visualisation de données comme un nuage de points **(voir Figure 15)**, on réalise une synthèse spatiale de l'hétérogène qui fait apparaître la relation de dépendance entre les variables dont on avait l'intuition : le nombre d'amis augmente avec le temps passé sur le site. Si l'on reproduit l'exercice non pas sur un utilisateur, mais un échantillon ou la totalité des utilisateurs du réseau social, on va pouvoir faire émerger une régularité, une norme relative à ces variables, et par rapport à laquelle on pourra explorer des variations. Par sa fonction de synthèse, la visualisation joue un rôle similaire à la modélisation statistique dans les sciences de la nature, un rôle de représentation simplifiée de l'objet qui devient intelligible et manipulable par son intermédiaire.

Notons tout d'abord que la visualisation de données est une forme de représentation graphique qui présente un caractère mécanique : contrairement à un dessin, une illustration, un schéma, elle produit un résultat précis à partir de la forme choisie et des données représentées. En ce sens, sa production n'est pas soumise à un travail interprétatif : deux personnes visant à représenter les mêmes données, par exemple une répartition hommes/femmes sous la forme d'un diagramme circulaire, obtiendront les mêmes valeurs d'angles pour représenter les secteurs circulaires correspondants au nombre d'hommes et de femmes ; il n'y a qu'une formule pour transformer les valeurs quantitatives correspondantes en surfaces, étant donné la forme et le contenu de la visualisation. Du fait de ce

caractère nécessaire, une visualisation peut être **générée comme un calcul**, par une machine (ou plus précisément, un logiciel, ou un langage de programmation doté d'un moteur graphique).

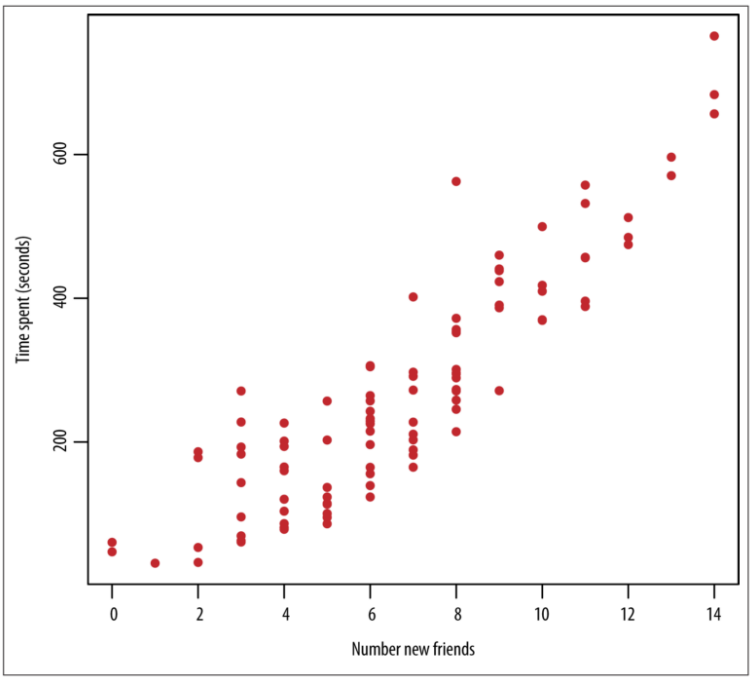

***Figure 15**. « Nuage de points » représentant les variables « temps passé sur le site » et « nombre d'amis » d'un utilisateur. La corrélation positive entre les variables (lorsque l'une augmente, l'autre augmente également) est aisément lisible.*

Historiquement, l'apparition de logiciels de visualisation a largement facilité et répandu l'usage de la visualisation de données. Bertin (1970) souligne ainsi que contrairement à d'autres formes sémiotiques comme les images figuratives (ou non), qui peuvent être soumises à une multitude d'interprétations relevant de l'esthétique, de l'histoire, de la culture, etc., la visualisation de données (qu'il appelle « la graphique »[49]) est monosémique. La signification des signes graphiques s'y déduit mécaniquement comme dans le langage mathématique. Néanmoins, contrairement à ce langage qui suppose une lecture linéaire, inscrite dans une durée temporelle, monodimensionnelle[50], « la graphique » s'inscrit dans l'espace formé par les deux dimensions du plan et la variation chromatique, ce qui lui donne un caractère tridimensionnel. Par ailleurs, elle relève d'une perception immédiate, alors que la perception est seulement un préliminaire à la lecture dans le cas d'un langage linéaire. Dans la visualisation de données, l'espace prend un statut représentationnel où il n'est pas seulement autodescriptif : les positions, les distances, les couleurs, deviennent un « "système de signes", complet, indépendant, et possédant ses lois propres, c'est-à-dire sa "sémiologie" » *(Ibid.)*.

[49] Contrairement à Jack Goody chez qui la liste, le tableau, sont des formes graphiques, « la graphique » chez Bertin désigne uniquement la visualisation de données.
[50] Cette affirmation de Bertin est d'ailleurs contestable, dans la mesure où la lecture de formules mathématiques complexes n'est pas purement linéaire : en pratique, on évalue tour à tour sa forme générale, ses différents éléments, entre lesquels on navigue jusqu'à être en mesure de procéder à une lecture syntaxique de la formule dans son ensemble.

Bertin définit huit **variables visuelles** pour toute visualisation de données :

- la position horizontale et verticale de l'élément ;
- sa forme (rectangle, carré, triangle, forme libre, etc.) ;
- son orientation ;
- sa taille ;
- sa « valeur » (c'est-à-dire sa couleur par rapport à une palette comprenant par exemple un rouge, un jaune, un bleu) ;
- sa « couleur » (plus exactement, l'intensité de sa couleur, rouge presque blanc, rouge foncé, etc.) ;
- son « grain » (c'est-à-dire, dans le cas d'un coloriage de la surface par trame de ligne (hachures), l'épaisseur de ces rayures, ou pour un coloriage par trame de points, la taille de ces points).

Toute visualisation de données, qu'il s'agisse d'un diagramme, d'une carte géographique ou d'un graphe représentant un réseau, peut être intégralement décrite en fonction de ces variables. La signification de ces variables est partiellement **conventionnelle :** par exemple, sur une carte géographique, on représente les fleuves par des traits bleus, les densités de population par des ronds de taille variable ou des surfaces dont l'intensité de couleur varie, etc. Néanmoins, rien n'empêche le géographe de représenter les fleuves en rouge et les mers en vert. Du point de vue de la sémiologie graphique, ces choix sont valides dès lors qu'ils sont appliqués de manière cohérente, et explicités, par exemple par une légende.

Si ces choix sont valides, ils ne sont en revanche pas forcément souhaitables. En effet, bien que la signification des éléments graphiques procède de la déduction et non de l'interprétation, l'analyste dispose d'une certaine liberté dans l'exécution, et doit réaliser un certain nombre de choix. Les logiciels de visualisation tendent à orienter ces choix en proposant des paramètres « par défaut », que l'on ajuste ensuite au lieu de devoir tout définir préalablement. Un tracé à la main, ou dans un logiciel de traitement d'image, obligera en revanche à anticiper tous ces choix (ou du moins une première version de ces choix) avant de les mettre en œuvre, mais tout en privant le graphiste du caractère largement automatisé de la génération d'une visualisation de données par un logiciel. Même dans ces conditions, les paramètres par défaut ne sont pas toujours, voire sont rarement les meilleurs, d'un point de vue sémiologique comme esthétique.

Cette liberté dans l'exécution de la visualisation porte à la fois sur le type de visualisation choisi, les variables visuelles, ou encore les éléments d'explication choisis (titres, légendes, etc.). Le support numérique ajoute également une dimension d'interactivité, par laquelle l'utilisateur peut filtrer, sélectionner, zoomer ou dézoomer, etc. Une répartition peut être représentée par un diagramme en bâton ou circulaire ; l'évolution d'une quantité en fonction du temps peut être représentée par une

ligne ou par un histogramme. Les cartes géographiques sont générées à partir d'une projection spécifique du globe terrestre, comme la projection équirectangulaire ou la projection de Mercator, qui transforme la surface du globe en plan. Les cartogrammes sont des types de cartes où l'espace géographique est déformé en fonction des quantités associées à une variable, par exemple une carte de l'Europe où la surface des pays augmente en fonction de leur population. Toutes ces variations au sein d'une même catégorie de visualisation constituent l'**espace de liberté** de l'analyste qui visualise des données.

Cet espace de liberté dans les pratiques est celui où émerge l'**art de la visualisation de données**. C'est un art dans le plein sens du terme car il requiert un savoir-faire technique tout en s'appréciant également sous des critères esthétiques. À partir de significations établies mécaniquement, l'analyste dispose d'une marge de manœuvre au sein de laquelle il peut obscurcir ou au contraire rendre clair le sens qui se dégage de la visualisation. Dans cet espace de liberté, certaines pratiques sont considérées comme bonnes ou mauvaises : ainsi, tronquer un axe pour atténuer une différence entre plusieurs barres d'un histogramme est perçu comme un choix malhonnête ; l'utilisation d'ombres, de déformations, d'effets 3D, est jugée nuisible. De manière générale, les longueurs sont préférables aux surfaces, car la traduction d'une quantité en longueur est linéaire, tandis que sa traduction en surface est au moins exponentielle. Les conventions de visualisation que nous évoquions ci-dessus font partie également des éléments qui facilitent la lecture. Des fleuves représentés en bleu n'ont généralement pas besoin d'être légendés pour être immédiatement compréhensibles, alors qu'une autre couleur peut créer de la perplexité. Le fait qu'un géographe convenable les reconnaîtra immédiatement non pas *via* leur couleur, mais *via* leur forme et leur position sur la carte, ne justifie pas l'adoption non motivée ni justifiée de choix non conventionnels.

Dans les domaines des interfaces hommes-machines et de l'ergonomie, qui s'intéressent à la visualisation de données et sont issus notamment des sciences cognitives, une visualisation peut être évaluée sur un **critère de lisibilité** mesurable. L'intelligibilité des visualisations de données fait ainsi l'objet de travaux depuis une quarantaine d'années, soit peu de temps après l'apparition de la visualisation de données par ordinateur. La particularité de ces travaux est qu'ils ne s'intéressent pas à la compréhension des visualisations d'un point de vue cognitif mais perceptif, distinguant ce qui relève de la pensée et de la vision[51]. Ainsi, Cleveland et McGill (1985) étudient la « perception graphique élémentaire » ou la « vision pré-attentive » plutôt que la cognition :

> Visual decoding as we define it for elementary graphical-perception tasks is what Julesz calls preattentive vision (6): the instantaneous perception of the visual field that comes without apparent mental effort. We also perform cognitive tasks such as reading scale information, but much of the power of graphs - and what

[51] De ce fait, ces travaux s'inscrivent dans un cadre théorique où la vision n'est pas une tâche cognitive, ou pas exclusivement, ce qui est par ailleurs sujet à débat notamment lorsque l'on se situe dans un cadre enactiviste.

> distinguishes them from tables - comes from the ability of our preattentive visual system to detect geometric patterns and assess magnitudes. We have examined preattentive processes rather than cognition.

Les dispositifs de mesure sont ceux de la psychologie expérimentale : un échantillon d'individus est soumis à un dispositif expérimental suivant lequel il doit réaliser certaines tâches et répondre à certaines questions. Par exemple, on place les individus devant un diagramme et leur demande de répondre à des questions telles que « quelle est la plus grande mesure ? », « l'élément A est-il plus grand que l'élément B ? », « l'élément B a-t-il la même taille que l'élément A ? », ou encore « quel est le ratio de l'élément A sur l'élément B ? ». Sont mesurés leur temps de réponse aux questions et leur taux d'erreur, qui permettent alors de comparer quantitativement différents types de visualisations, ou différentes variantes d'une même visualisation. Ce type de travaux met en évidence les difficultés de lecture liées notamment aux diagrammes circulaires et aux graphiques avec un effet 3D (Stevens 1981; Marr 1982) et déjà évoquées aux chapitres V et VI.

Ces travaux s'intéressent plus spécifiquement à l'effet des variations de taille, de forme et de position. Les autres variables visuelles sont également des axes d'amélioration de la lisibilité du graphique. Des contrastes suffisants entre couleurs sont recommandés pour tout travail de création visuelle. Sur ce point, le W3C, organisme de standardisation du web, s'appuie pour ses recommandations sur des travaux en ophtalmologie concernant la lisibilité chez les personnes malvoyantes[52]. Il s'agit à la fois de veiller aux contrastes :

- entre teintes (un violet opposé à sa couleur complémentaire, le jaune, sera plus contrasté qu'un violet opposé à une couleur voisine comme le rouge),
- en termes de saturation (un violet vif par rapport à un violet tirant vers le gris),
- et en termes de luminosité (un violet foncé, presque noir, par rapport à un violet clair, presque blanc).

Par ailleurs l'utilisation de couleurs distinguables suivant les différentes formes de daltonisme fait partie des bonnes pratiques de visualisation de données.

Enfin, un autre axe de lisibilité relève de l'établissement d'une **économie graphique** qui élimine les éléments visuels inutiles. Edward Tufte, que nous avons également déjà cité plusieurs fois, propose ainsi dans *The Visual Display of Quantitative Information* (2008) une approche dogmatique de cette économie, dans laquelle il faut absolument maximiser la quantité d'encre (ou de pixels non transparents) présentant de l'information quantitative. En ses termes,

> Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.

[52] https://www.w3.org/TR/UNDERSTANDING-WCAG20/visual-audio-contrast7.html (consulté le 16 novembre 2017)

Idéalement, chaque goutte d'encre (ou chaque pixel) utilisée pour la visualisation doit servir à représenter une quantité, de manière à ce que rien ne puisse être effacé sans perdre de l'information : il faut optimiser le **ratio donnée-encre** (*data-ink ratio*). Le reste est du déchet de graphique *(chartjunk)* qui doit être éliminé. Il arrive par ailleurs que les variables graphiques soient si mal exploitées que le lecteur se retrouve contraint de lire les libellés affichant les valeurs numériques de la série de données. Dans ces conditions, la visualisation est inutile, puisque le lecteur réalise les mêmes opérations mentales que devant une ou plusieurs listes de chiffres, et retourne à une lecture linéaire. L'enjeu de la visualisation de données est précisément d'éviter, ou de réduire au maximum, ce besoin de revenir à une lecture linéaire, en rendant possible une lecture spatiale et chromatique non linéaire.

Pour illustrer cette exigence d'**excellence graphique**, il prend plusieurs exemples de visualisation tirés de la presse ou de livres et les dessine en suivant ses préceptes. La **Figure 16** et la **Figure 17** illustrent ainsi sa révision de la représentation de l'évolution du budget de l'État de New York entre 1966 et 1967.

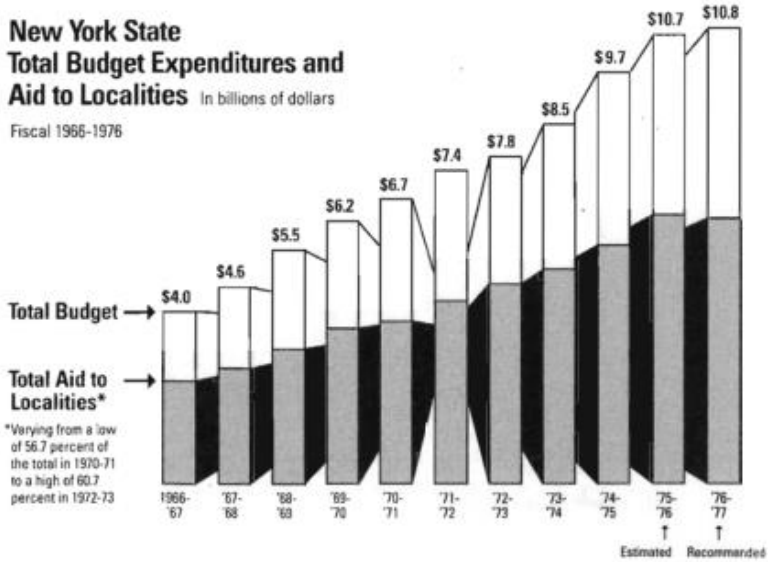

***Figure 16**. Graphique tiré du* New York Times *du 1e février 1976, présenté par Edward Tufte* (*Ibid.*) *comme un exemple malhonnête de visualisation de données visant à suggérer que le budget prévu pour 1977 est supérieur à toutes les années précédentes.*

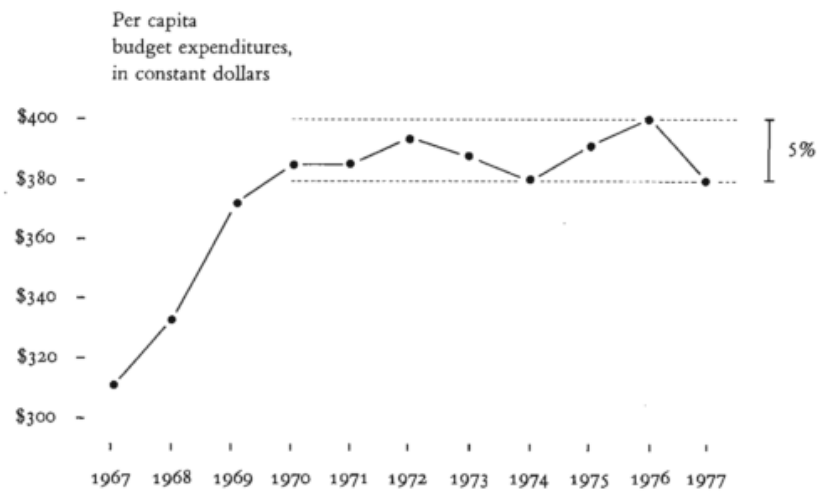

***Figure 17**. Le même graphique que la **Figure 16**, redessiné par Tufte suivant ses préceptes. De nombreux éléments ont été supprimés dont les effets d'ombres, les bordures, les annotations, certaines échelles. Les valeurs indiquées ont également été modifiées. Par ailleurs, la population de l'État de New York a augmenté et le dollar a subi une inflation substantielle dans l'intervalle représenté. De ce fait, une mesure différente est utilisée pour montrer l'évolution du budget en prenant en compte ces facteurs : la dépense par habitant en dollars constants. Le ratio consacré aux aides locales, qui exprime une autre idée, a été éliminé ; enfin, une annotation met en évidence que la variable exprimée a diminué de 5% entre 1976 et 1977, contrairement à ce que suggérait le graphique original.*

Les critères de lisibilité et de ratio donnée/encre sont des critères d'application générale, quelle que soit l'idée exprimée par un graphique. Ils constituent des normes ou des recommandations pour une sémiologie graphique efficace. Néanmoins, comme on le voit avec l'exemple de la **Figure 17**, la question de déterminer *comment* on montre l'information conduit rapidement à une nécessité de réviser *ce que* l'on montre. Ainsi, dans la révision de Tufte, la variable représentée elle-même change : ce n'est plus le budget en valeur absolue, mais la dépense par habitant (pour neutraliser l'impact de l'augmentation de la population) et en dollars constants (pour neutraliser l'influence de l'inflation). Sans cette modification, les valeurs numériques visualisées représentent en réalité des informations différentes, l'augmentation de la population et l'inflation ; la modification permet de recentrer la visualisation sur l'expression d'une seule idée. Le besoin de mise en évidence amène à retravailler sur les éléments statistiques matérialisés par la visualisation de données, et à mobiliser des connaissances du domaine et des informations externes. Ainsi, en l'occurrence, des données supplémentaires ont été utilisées pour calculer la nouvelle variable :

- des connaissances qualitatives, comme le fait que le budget de l'État est corrélé à la taille de la population, et le concept d'inflation ;
- des connaissances quantitatives, à savoir la démographie de l'État de New-York entre 1967 et 1977, et la mesure de l'inflation sur cette même période.

Enfin, une compétence supplémentaire est requise, de nature arithmétique : la capacité à calculer une nouvelle variable en combinant de la bonne façon la série de données initiales et les connaissances quantitatives ajoutées.

En phase exploratoire comme dans la préparation de la restitution, l'amélioration de la visualisation permet de mieux donner à voir une idée. Pour l'analyste en phase exploratoire, il s'agit de discerner avec une acuité croissante l'information qui se dégage, c'est-à-dire de la séparer des autres informations matérialisées par la visualisation, et du bruit statistique qui peut apparaître ; il y a ainsi découverte ou confirmation progressive d'une idée à mesure que celle-ci est séparée de la multitude d'informations présentes dans les données. En phase de restitution, il s'agit de mettre en évidence l'idée de la manière la moins ambiguë possible, d'empêcher toute confusion : quelles que soient les conditions de lecture, les connaissances et la culture du lecteur, la visualisation doit avoir la même signification pour tous les lecteurs qu'elle vise. Cette remarque est importante car si on peut souhaiter qu'une visualisation soit universelle, on peut également la constituer pour un contexte de restitution précis : ainsi, une carte dessinée pour une convention de géographes ne s'inscrira pas dans les mêmes conventions de représentation qu'une carte destinée à un article de presse et à une diffusion grand public, par exemple.

Il est toujours possible d'éviter la confusion dans l'interprétation en écrivant par ailleurs l'idée que l'on souhaite exprimer par la visualisation. Néanmoins, l'écriture « en bon français » de l'idée à exprimer, ne se substitue pas à la visualisation, à supposer qu'elle lui soit équivalente. En effet, la visualisation joue par ailleurs un rôle de preuve : écrire que le budget n'a pas augmenté en 1977, ou qu'il a diminué de 5% par rapport à 1976 (deux formulations textuelles de la même idée), n'est pas la même chose que le montrer avec une visualisation de données qui présente une justification quantitative de l'information. Pour que la substitution soit complète sur une visualisation simple comme une courbe d'évolution temporelle, il faudrait en réalité expliciter toutes les valeurs numériques, leurs différences en valeurs absolue et en pourcentage, qui sont autant d'informations visuellement accessibles sans être explicitées.

Même si cette explicitation textuelle ne présente pas un grand intérêt, elle reste envisageable dans le cas d'une visualisation simple comme une courbe de tendances. Les visualisations les plus courantes présentent un niveau de simplicité équivalent ; elles sont presque toutes construites autour de trois éléments que sont le bâton, la courbe et le cercle. Un diagramme circulaire peut d'ailleurs être redessiné comme un diagramme en bâton ; une courbe de tendances, comme un histogramme[53]. Les

[53] Soulignons ici qu'un diagramme en bâton n'est pas la même chose qu'un histogramme : le premier représente les quantités associées aux valeurs prises par une variable qualitative (par exemple, des quantités associées à « femmes » et « hommes ») tandis que le second représente une variable continue discrétisée (par exemple, l'âge, réparti en tranches). La convention pour les distinguer, en dehors de leur signification, est d'espacer les bâtons dans le cas du diagramme en bâton et non pour l'histogramme.

visualisations reposant sur ces trois éléments peuvent généralement être décrites textuellement sans grande difficulté ; c'est loin d'être le cas, en revanche, de la plupart des autres types de visualisations.

Le monde universitaire affectionne en effet comme élément de base le point, qui sera spatialisé suivant deux coordonnées représentant deux variables, comme sur la **Figure 15** proposée par Rachel Schutt et Cathy O'Neil. Un nuage de points est largement plus complexe à décrire textuellement. Cette visualisation montre en effet :

- la position de chaque point ;
- sa distance aux autres points ;
- la forme générale du nuage (est-ce plutôt une ligne, un rond, plusieurs ronds ?) ;
- la densité du nuage (les points sont-ils resserrés autour d'une ligne ?) ;
- les points « aberrants » (éloignés de la forme générale) ;
- etc.

Du fait que le nuage de points matérialise aisément plusieurs dizaines ou centaines d'éléments, une description textuelle deviendrait infiniment longue. Cette remarque est valable également pour les cartes géographiques, dont la traduction textuelle serait interminable aussi bien qu'inintelligible. Dans ces exemples, la complexité de la description textuelle permet de rendre compte de la quantité d'information matérialisée dans une visualisation comme le nuage de points ou la carte. Or, le contexte des *big data* favorise ces représentations plus riches, en particulier en phase exploratoire. Le plan factoriel de l'analyse de correspondances, la visualisation de graphe, les matrices de cooccurrences, sont autant d'éléments qui ne servent pas tant à représenter une idée qu'à rechercher des idées intéressantes, ou à matérialiser un ensemble d'idées. Ce type de visualisation s'apprécie et s'évalue sous des critères spécifiques, distincts des normes explicitées par Bertin ou Tufte.

2. La cartographie de données

Nous proposons d'appeler ces visualisations plus riches et complexes des ***cartographies***, c'est-à-dire des représentations systématiques et exhaustives d'un objet spécifique, qui dégagent une vue globale et synthétique de cet objet tout en permettant de rechercher des idées plus précises :

- D'une part, une cartographie a un aspect **systématique** du fait qu'elle déroule l'ensemble des combinaisons possibles, généralement entre deux ou plusieurs variables caractérisant l'objet : de ce fait, elle constitue une représentation globale de l'objet. Naturellement, la combinatoire porte sur la représentation statistique dont on dispose et non sur l'objet lui-même, qui n'est pas donc épuisé par sa représentation statistique ou visuelle ;
- D'autre part, cette représentation globale constitue un espace que l'on peut parcourir, à la recherche d'**indices**. Le type de regard que l'on porte n'est pas un regard global, synthétique,

mais un regard analytique qui décompose l'espace en zones susceptibles de matérialiser une idée intéressante.

Ainsi, une carte de chaleur (**Figure 18**) est un genre de cartographie qui épuise la combinatoire entre deux variables tout en matérialisant deux mesures quantitatives : le volume associé à chaque combinaison, et un score de surreprésentation qui exprime une relation entre deux éléments. Une analyse factorielle de correspondances explore systématiquement, en les simplifiant, les relations entre plusieurs variables. Une visualisation de graphe matérialise et spatialise l'ensemble des relations pondérées entre éléments. Dans l'ensemble, ces cartographies s'appuient sur trois types d'information : des mesures ou quantités, des relations (en particulier pour la visualisation de graphes) et des coordonnées (pour les cartes géographiques).

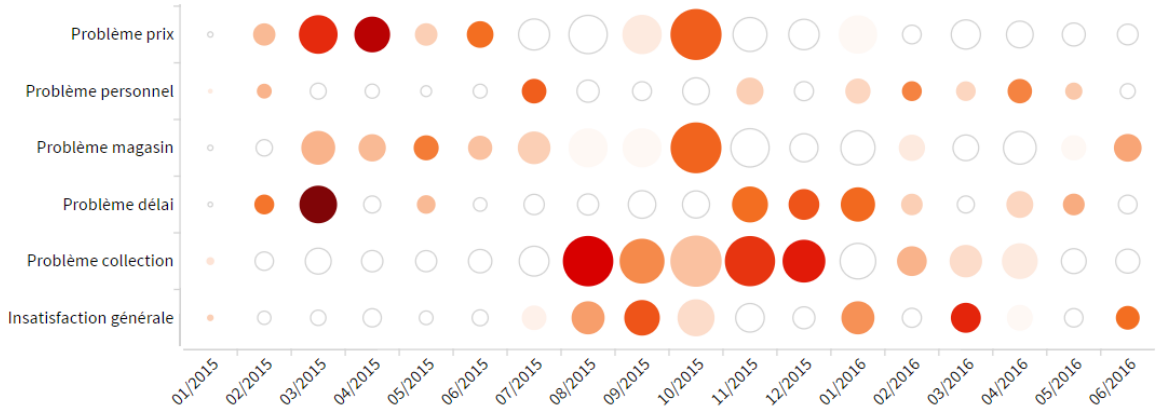

Figure 18. Heatmap *réalisée avec le logiciel de Proxem, représentant les types de problèmes identifiés sur un ensemble de verbatims clients sur des enseignes de prêt-à-porter, croisés avec l'évolution du volume dans le temps. La taille des cercles représente le volume de verbatims correspondant à chaque combinaison, tandis que l'intensité de la couleur représente un score de surreprésentation. Plusieurs zones attirent l'attention, comme par exemple les problèmes de délai en mars 2015, les problèmes sur la collection entre août et décembre 2015, ou encore l'insatisfaction générale en mars 2016.*

La sémiologie des cartographies est distincte de celle des visualisations plus simples analysées par Tufte : elles ne servent pas à représenter une idée spécifique mais un **espace représentationnel**, dont les propriétés relèvent bien d'une sémiologie graphique (les positions, les couleurs, les formes, ont une signification programmée), et dans lequel l'interprétation fait émerger diverses idées potentiellement intéressantes. De ce fait, les notions d'excellence graphique avancées par Tufte ne s'appliquent pas, ou seulement partiellement, à ce type de visualisation, en particulier lorsqu'elles sont des objets intermédiaires de travail. Le sens d'une cartographie a rarement un caractère d'évidence et nécessite d'être élaboré progressivement à travers l'exploration de l'espace cartographique. La première version d'une cartographie, générée avec les paramètres par défaut du logiciel, permet à l'analyste de vérifier que la représentation générée est bien celle qui était recherchée, mais pas encore à en tirer une signification. Elle suscite généralement un sentiment de désorientation devant l'abondance de points représentés.

L'effort herméneutique peut en effet être perturbé par une surcharge esthétique. Dans un article sur la visualisation de données dans le cadre des *big data*, Anthony McCosker et Rowan Wilken (2014) comparent la désorientation éprouvée devant une visualisation à l'expérience du **sublime** mathématique chez Kant, lui-même inspiré du sublime chez Burke. Burke s'intéresse en effet aux larges objets, qui relèvent de quantités si grandes que l'imagination ne leur trouve pas de fin, de sorte que cette dernière, ne parvenant à percevoir les frontières de ces choses, les croit infinies ; de ce fait, elles produisent les mêmes effets que si elles l'étaient effectivement.[54] Dans la *Critique de la faculté de juger* (1995), Kant définit le sentiment de sublime comme :

> un sentiment de déplaisir provenant de l'inadéquation qui dans l'évaluation esthétique de la grandeur, caractérise l'imagination à l'égard de l'évaluation de la raison - et il s'y trouve en même temps un plaisir suscité par l'accord entre précisément ce jugement sur l'inadéquation du plus grand pouvoir sensible et les Idées de la raison, en tant que c'est cependant pour nous une loi que de faire effort pour atteindre ces Idées.

En d'autres termes, le sublime suscite un sentiment simultané de plaisir et de déplaisir. Le sentiment de déplaisir vient de l'incapacité de l'imagination (c'est-à-dire ici, de la capacité à appréhender une perception) à rendre raison de la grandeur perçue. Le sentiment de plaisir vient du fait qu'en étant rendue sensible, cette incapacité de l'imagination révèle combien ces grandeurs sensibles, si grandes soient-elles, sont petites en comparaison des Idées de la raison, auxquelles il nous faut aspirer chez Kant ; cette aspiration est la source du sentiment de plaisir.

La perception d'une grande quantité d'objets, ou de grands espaces, révèle notre incapacité à les appréhender de manière synthétique, comme une unité : l'emprise esthétique de l'objet prend le dessus sur son appréhension épistémique. Nous percevons successivement chaque élément sans pouvoir en faire une synthèse porteuse de sens. En ces termes, une visualisation de type cartographie peut ainsi susciter un sentiment de sublime : la quantité d'éléments élémentaires et de variation graphique sur ces éléments, est une grandeur que la perception ne peut appréhender comme une unité. Néanmoins, contrairement à ce que suggèrent McCosker et Wilken, toutes les visualisations de données générées à partir de *big data* ne provoquent pas ce sentiment : la synthèse graphique de grands volumes que réalise un diagramme en bâton ou une courbe de tendances permet au contraire d'appréhender synthétiquement toutes les données représentées et de les ramener à la fois dans les frontières de la perception sensible et de l'appréhension intellectuelle. Le problème posé par les visualisations de type cartographie est qu'au lieu d'éliminer la désorientation suscitée par l'abondance de données, ce « déluge » qui revenait comme un *leitmotiv* dans les discours sur les *big data*, elles reproduisent cette abondance, ce qui doit conduire l'analyste à travailler à l'intérieur de celle-ci pour recomposer du sens.

[54] 'some large objects are so continued to any indefinite number, that the imagination meets no check [and thus, by] not being able to perceive the bounds of many things, they seem to be infinite, and they produce the same effects as if they were really so' (citation de Burke tirée de McCosker & Wilken, 2014)

De ce fait, là où la visualisation simple résulte d'un processus qui vise à mettre en évidence quelque chose de quasi-factuel, qui ne laisse pas de place à l'interprétation, la cartographie nous renvoie au paradigme indiciaire et à l'exploration de pistes, d'indices, dans un espace désormais graphique et non plus seulement statistique, dans lequel il s'agit de parvenir à s'orienter et à se repérer. Ce type de visualisation n'est pas le résultat d'une exploration, mais ce qui sert à explorer (outil), et ce qu'il y a à explorer (objet). Il n'est pas question d'appréhender la cartographie dans sa totalité, sans quoi on se heurte à un sentiment esthétique de sublime qui empêche l'activité épistémique, mais de regarder *à l'intérieur* de l'espace cartographique, à la recherche d'indices.

Ce type de visualisation appartient ainsi résolument à la phase exploratoire. On ne peut l'envisager comme forme de restitution qu'au terme d'un **travail d'ajustement** qui vise à faciliter l'interprétation au moment de la restitution. Ce **perfectionnement** est en soi une forme d'exploration qui fait intervenir à la fois le sens des données, la technique de cartographie et le rendu visuel. Il ramène, contrairement à la cartographie initiale, à la recherche d'excellence graphique de Tufte : ce perfectionnement va réduire la quantité ou la visibilité des éléments graphiques qui ne servent pas à exprimer une idée. Il vise également à faire disparaître, ou du moins à diminuer, le sentiment de sublime liée à la cartographie initiale, pour ramener sa dimension esthétique à une fonction herméneutique et rhétorique : herméneutique, car l'interprétation est plus facile car elle se fait sur un objet esthétiquement plaisant, et rhétorique, car son caractère esthétique séduit l'observateur avant d'avoir à le convaincre. Les modalités de cette phase d'ajustement ou de perfectionnement dépendent du mode de visualisation :

- Dans les *heatmaps* de Proxem, il s'agit de déterminer et d'ajuster les catégories représentées, l'ordre dans lequel les valeurs apparaissent, le score de surreprésentation utilisé, mais aussi le ratio longueur/largeur de la visualisation ou la taille des libellés dans une optique de lisibilité.
- Dans le cas de l'analyse factorielle de correspondances, le travail consiste à évaluer les différents axes représentés, à masquer certains points, à recadrer sur des portions intéressantes du plan factoriel, mais aussi à éviter la superposition de libellés, ajuster les couleurs, etc.
- Dans le cas de la visualisation de graphe, les logiciels comme Gephi sont presque entièrement dédiés à la fabrication de la représentation visuelle, avec deux phases : celle qui va placer les nœuds dans l'espace, définir quelles variables régissent la taille et la couleur des nœuds, et une seconde phase plus cosmétique où sont ajustées les épaisseurs, les bordures, la typographie, etc.

Ces phases d'ajustement, voire de tâtonnement, sont emblématiques de l'analyse exploratoire de données au sens où elles permettent à la fois de se familiariser avec les données, d'en dégager progressivement une ou plusieurs interprétations, et de préparer une forme de restitution. De cette

manière, une cartographie peut devenir une forme de restitution à condition d'être travaillée en fonction d'une certaine interprétation. C'est un travail technique qui s'emploie par ailleurs à enlever à la cartographie son caractère de sublime, de manière à ramener la dimension esthétique de la visualisation à une fonction rhétorique au service de son intelligibilité. Cette interprétation peut être matérialisée de plusieurs façons (qui ne sont pas mutuellement exclusives) :

- directement dans la visualisation, en gommant ou en masquant les éléments qui génèrent du bruit par rapport à la lecture recherchée ;
- en surchargeant la visualisation avec d'autres éléments graphiques (des flèches, des zones entourées ou surlignées) qui attirent le regardent sur une zone particulière de la visualisation ;
- par explicitation textuelle *via* des titres, des libellés, des légendes, qui peuvent être combinés avec les deux points précédents.

Un exemple combinant ces trois points serait, dans le cas d'une visualisation de graphe, de griser tous les nœuds du graphe sauf ceux sur lesquels on veut attirer l'attention (transformation directe de la visualisation), de pointer sur la zone concernée avec une flèche et de placer un commentaire textuel au bout de la flèche. De cette façon, la cartographie garde sa capacité à faire preuve dans la mesure où elle contextualise l'élément mis en avant et permet de le replacer dans son ensemble, tout en facilitant la lecture et la compréhension.

Cette phase d'ajustement est le propre du contexte exploratoire, mais c'est aussi celle qui prépare la forme de restitution. Dans les termes de Reichenbach, c'est un instrument pour la découverte mais aussi pour la justification. En explicitant l'interprétation par des éléments textuels et graphiques, l'analyste prépare la visualisation à être partagée à d'autres. Dans le cas d'une configuration en récit en particulier, la visualisation de données est une partie intégrante dudit récit. Ce constat est particulièrement vrai des formes de type « narration par la donnée » *(data storytelling)* qui se donnent explicitement pour finalité de ménager la plus grande place à la visualisation, au détriment d'un récit textuel. Dans ce type de format, si l'on considère que le récit est la forme globale, la visualisation est à la fois son contenu, les parties d'un tout, et une forme spécifique, visuelle, imbriquée à l'intérieur de la forme générale « récit ».

Notons ici que la visualisation de données occupe une place tout à fait distincte de celle de l'illustration dans les œuvres de fiction. Il ne s'agit pas de reproduire, d'illustrer, ce qui est déjà présenté dans le récit, mais de se substituer à ce qui aurait pu être une portion de texte. Comme on l'a vu en effet, l'efficacité narrative et la lisibilité de la visualisation de données sont supérieures à une restitution sous forme textuelle dans le contexte de l'analyse computationnelle de données. Les éléments textuels qui surchargent les visualisations ne sont pas des éléments de récit autonomes à proprement parler, mais des aides à l'interprétation qui complètent la visualisation et précisent son sens, à la façon dont une note de bas de page peut clarifier une idée développée. La visualisation de données n'a pas non

plus le même statut dans le *data storytelling* et dans une publication scientifique. Pour cette dernière, on remarque en effet que le fil du texte ne s'interrompt pas pour laisser la visualisation se substituer à lui temporairement : les graphiques proposés sont à côté du texte, signalés par des renvois. L'information exprimée par le graphique est répétée dans le fil de l'article, et souvent dans la légende. Le récit visuel ne se substitue pas au récit textuel, mais le complète, notamment en jouant un rôle de preuve.

3. La représentation comme preuve

En effet, la visualisation de données n'est pas un dessin tracé librement en fonction d'une intention ou d'une idée qu'on souhaiterait représenter, mais un résultat d'abord généré par un programme, puis transformé. Ce caractère **initialement génératif** la fait apparaître comme la conséquence mécanique des données, dont elle hérite les propriétés épistémiques. De ce fait, elle ne sert pas seulement à exprimer efficacement une idée, mais à la justifier. La valeur épistémique des données, qui leur était attribuée au moment de leur constitution ou agrégation, et qui était maintenue après leur manipulation computationnelle, est reportée dans leur visualisation du fait de ce caractère mécanique. C'est parce que ces données ne sont pas « naturelles », purement empiriques, mais théoriquement chargées, que cette charge théorique peut se propager à travers les différentes manipulations qu'elles subissent.

Dans le cas de Proxem par exemple, la valeur attribuée aux données relève avant tout d'un **contrat épistémique** avec le client : les données à traiter sont considérées comme ayant toujours déjà, avant leur traitement, une certaine validité, une certaine capacité à représenter l'objet que l'on cherche à étudier. Ainsi, la charge de la preuve de validité des données n'appartient pas à Proxem, mais aux agents qui ont constitué le jeu de données. Dans le cas des *digital methods*, c'est l'analyste qui, pour constituer les données, définit une **logique d'acquisition** motivée par un projet épistémique. Comme on l'a vu au chapitre IV, il est motivé à la fois par une finalité de représentation de phénomènes sociaux, par un état d'esprit empiriste et expérimental fonctionnant sur le mode de l'essai-erreur, et par une vision critique de cette démarche que l'on peut voir comme une **justification par auto-falsification**. En effet, si, en essayant d'invalider sa logique de constitution de corpus, l'analyste n'y parvient pas, cet échec apparaît comme un élément de validation de la démarche. Cette falsification peut également mettre en évidence des limites qui sont autant de déterminations de la portée des données, du domaine de connaissances possibles qu'elles sont susceptibles de couvrir.

Le geste théoriquement chargé de constitution du jeu de données est un point de départ qui confère un potentiel épistémique à ces données. Ce potentiel s'actualise par les manipulations que subissent les données et notamment leur traitement graphique. Au cours de ces différentes étapes, la valeur épistémique de la donnée est transférée mécaniquement d'une étape à l'autre, tout en faisant l'objet

d'un contrôle épistémique assuré par la posture interprétative. À ce titre, le traitement de données relève de plusieurs formes d'objectivité : l'objectivité mécanique et le jugement exercé.

Comme l'ont montré Lorraine Daston et Peter Galison dans leur ouvrage *Objectivité* (2007), il existe plusieurs formes d'objectivité, historiquement constituées. L'objectivité n'y apparaît pas tant comme une propriété intrinsèque des savoirs scientifiques que comme un ethos auquel le chercheur s'identifie dans sa pratique de la science. Parmi ces formes d'objectivité, nous en retiendrons deux qui correspondent à des modes de constitution et de justification des connaissances dans les sciences des données :

- L'**objectivité mécanique** est un ethos qui repose sur l'effacement de la subjectivité de l'individu, de son ego, de ses idiosyncrasies stylistiques en substituant à l'artisan humain un instrument « sans volonté ». Dans ce sens, la photographie fut perçue, au moment de son invention, comme plus objective que la peinture car elle était tenue pour capable de représenter la nature telle qu'elle est, contrairement à un artiste qui y imprègnera son style, ses affects, ses défauts, etc.
- L'objectivité comme effort actif fondé sur le **jugement exercé** d'un sujet doué de volonté, nécessaire pour produire une représentation fidèle à l'objet. C'est cette forme qui est à l'œuvre lorsque le médecin expérimenté est capable, à force de pratique, de détecter une pathologie en examinant une radio. Elle implique des notions d'intuition et d'interprétation qui viennent contredire ou compléter le caractère mécanique de l'image produite.

Dans la description générale de Daston et Galison ces deux formes d'objectivité ne caractérisent pas les observables eux-mêmes mais l'activité et la posture épistémique par rapport aux observables. Dans nos termes, on dira qu'elles ne concernent pas les données elles-mêmes mais les opérations réalisées sur elles. Ce ne sont pas les données qui sont objectives, mais les traitements. L'objectivité apparaît alors non pas comme un régime de validité des connaissances, où les connaissances sont valides car perçues comme objectives, mais comme une condition requise pour préserver cette validité. Un traitement objectif est un traitement qui **préserve** la valeur épistémique de la donnée. À ce titre, l'objectivité est construite en deux étapes :

- Les traitements computationnels, dont les techniques de fouille de données et d'apprentissage automatique, ainsi que la génération initiale des visualisations de données, sont des processus mécaniques qui préservent automatiquement la valeur épistémique de la donnée.
- L'interprétation et le perfectionnement des cartographies, mais aussi la préparation, le formatage et le nettoyage des données, ainsi que le paramétrage des algorithmes utilisés, sont des étapes qui nécessitent d'intervenir par-dessus les traitements mécaniques en vertu du jugement exercé.

La préparation des données est un bon exemple d'étape illustrant le fait que la seule objectivité mécanique ne suffit pas : en fonction du format initial des données, ne pas les retravailler serait une source de bruit, de résultats aberrants, voire rendrait impossible leur exploitation computationnelle. Il est donc nécessaire de faire intervenir un jugement exercé par le moyen duquel un certain nombre de choix sont effectués afin de rendre les données exploitables. Le même jugement exercé se retrouve dans le perfectionnement de la cartographie, qui permet de passer d'une visualisation inintelligible, perturbée par la désorientation du sublime, à une visualisation qui rend possible l'interprétation. L'objectivité du jugement exercé ne permet donc pas seulement de préserver la valeur épistémique des données, mais de la **révéler**. Elle résulte d'une herméneutique de la donnée qui nécessite la mobilisation de **connaissances externes** issues des sciences de la culture, mais aussi la maîtrise de la sémiotique correspondant à la technique de visualisation utilisée. Par exemple, le langage de la visualisation de graphes n'est pas le même que celui d'une analyse factorielle de correspondances. Ils reposent sur des ensembles de signes graphiques qui ont une signification programmée, à laquelle s'ajoute l'interprétation contextuelle de cette signification. Par exemple, la proximité géométrique s'interprète, en analyse factorielle de correspondance, comme une similarité entre éléments : c'est sa signification programmée. Cette proximité géométrique s'interprète ensuite en fonction des éléments représentés : il peut s'agir d'une proximité politique, musicale, historique, géographique, conceptuelle, etc. L'art de l'interprétant consiste précisément à contextualiser cette signification programmée, pour déterminer où elle est pertinente, mais aussi où elle relève de l'accident du fait d'un biais ou d'une aberration statistique.

Révéler la valeur épistémique de la donnée, c'est donc à la fois donner sens à certains signes, et écarter ceux qui n'en ont pas. Si cette valeur doit être révélée, c'est parce qu'elle n'est pas univoque ni évidente. En effet, il faut d'une part l'art du jugement exercé pour identifier les éléments de sens, et d'autre part, cette donnée doit être manipulée, manœuvrée, pour que son sens apparaisse. Les ajustements et perfectionnements effectués dans le cadre d'une visualisation de graphe ne sont pas des actions qui déforment une donnée brute pour la faire mentir, mais au contraire un ensemble de gestes qui donnent forme à la donnée pour qu'elle devienne intelligible. Comme on le voit à la **Figure 19**, un graphe « brut », sans ce travail d'ajustement est complètement illisible. Le même graphe (à droite), retravaillé, spatialisé, coloré, suggère des éléments de sens sans même que l'on sache ce qui est représenté. Chez Daston et Galison, le jugement exercé n'est pas considéré comme un point de vue subjectif qui pervertirait une image objective, qui transformerait la donnée brute en donnée dénaturée. Il est au contraire la source de l'objectivité de la représentation, ce qui lui donne son vrai sens. De même, dans la visualisation de données, comme dans toutes les phases d'interprétation des sciences des données, ce jugement exercé est ce qui rend la donnée lisible. La différence avec l'analyse de Daston et Galison sur le jugement exercé est qu'il s'agit aussi d'une *manipulation* exercée, un art de la transformation avisée de la représentation pour qu'elle devienne lisible et révèle son sens.

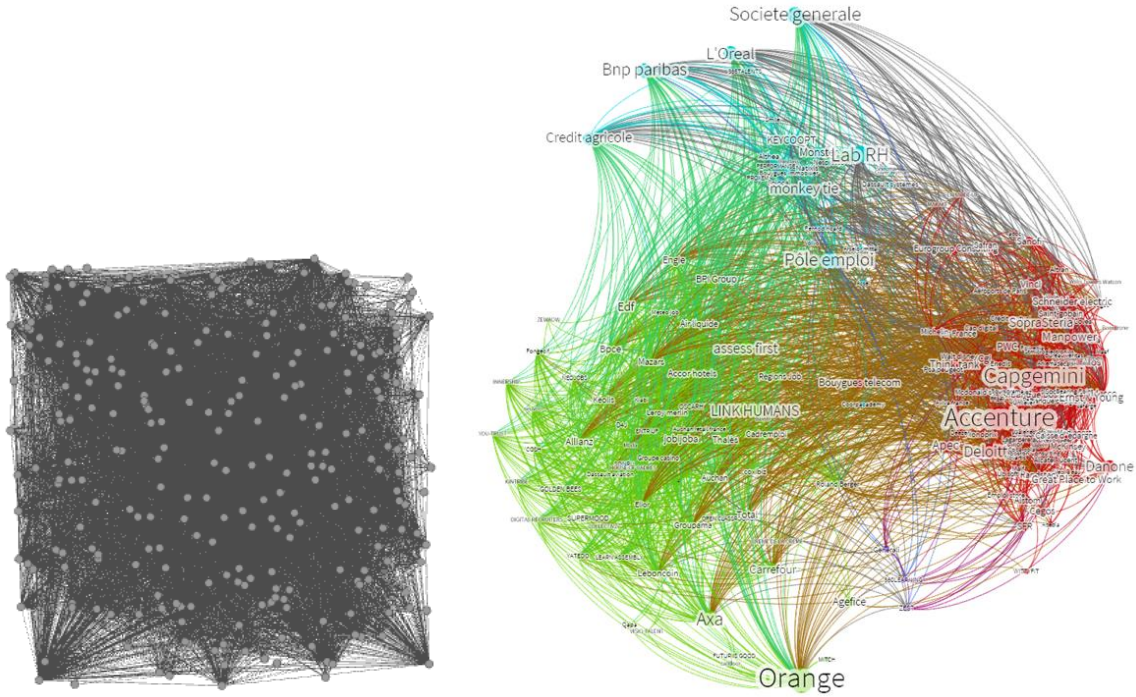

***Figure 19**. Le même graphe, représenté deux fois 1) sans ajustement ni spatialisation, tel qu'il apparaît lorsque le fichier est chargé dans le logiciel Gephi. 2) après le travail de « manipulation exercée » consistant à lui appliquer des actions de coloration, spatialisation, filtrage, mise en forme, etc. Le graphe « brut » est illisible tandis que le graphe retravaillé suggère des interprétations.*

Une représentation visuelle telle qu'une cartographie ne peut donc jouer son rôle de preuve que parce qu'elle résulte à la fois de procédures mécaniques qui préservent la valeur épistémique des données de départ, et d'une interprétation stabilisée par un jugement exercé. Ces opérations successives de manipulation, objectives car mécaniques ou avisées, préservent ou révèlent une valeur épistémique qui doit elle-même être justifiée par un discours falsifiable décrivant la logique d'acquisition de la donnée. Dans la configuration en récit, ce discours est le récit lui-même, qui décrit cette logique d'acquisition tout en retraçant les étapes de transformation de la donnée, dont la valeur est préservée ou révélée par ces étapes. Le récit remobilise le rôle exploratoire de la visualisation de données, à travers lequel l'analyste constitue l'interprétation de la donnée, tout en exploitant la capacité de la visualisation à faire preuve. La visualisation fait preuve parce qu'elle est en mesure de (re)présenter de manière lisible des significations constituées objectivement sur des données légitimement constituées, dès lors que cette objectivation et cette légitimation peuvent elles aussi pour leur part être explicitées par le récit.

Tout en ré-invoquant des étapes antérieures, nous avons ainsi montré comment la visualisation de données et son intégration dans la configuration en récit présentent, complètent et apportent une valeur de preuve aux résultats de la constitution et de l'analyse computationnelle de données. La visualisation fait preuve tandis que le récit justifie les assertions avancées par les éléments visuels, tout en les recontextualisant dans le contexte pragmatique du destinataire du récit.

Ces modalités de validation des connaissances, propres à la configuration en récit, ne se retrouvent pas ou peu dans la configuration en système, bien qu'elle mobilise elle aussi un langage visuel. Contrairement à la configuration en récit, ce langage, voire cette grammaire, n'y est pas celle de la visualisation de données, mais celui de l'interface telle qu'elle est conceptualisée et dessinée par le travail du designer.

4. Le langage visuel du design dans la configuration en système

La figure du designer est au moins aussi difficile à décrire que celle de l'ingénieur, avec l'obstacle supplémentaire qu'il s'agit d'une figure dont l'institutionnalisation est plus récente, démarrant autour des années 1970 (Petit et Deldicque 2017), et par rapport à laquelle on a donc moins de recul historique. Cette institutionnalisation est celle du design industriel, dont le rôle est de concevoir des produits manufacturés ; à ce titre on peut dire que cette figure incarne une partie des missions de l'ingénieur dans le contexte du génie industriel, et qu'elle émerge de celui-ci. Aujourd'hui, le design recouvre une multitude de problématiques plus spécifiquement traitées par le design d'information, le design d'interaction, le design d'interface, le design d'expérience, etc. Étant donné un objet à produire, on considérera le design comme 1) l'analyse des actions que l'objet doit rendre possibles et 2) la production d'un livrable (schème, maquette, prototype) qui procède d'une recherche à la fois esthétique et fonctionnelle, matérialisant la forme de l'objet. Cette dimension esthétique est centrale ; le design français utilise d'abord le terme d' « esthétique industrielle » (Petit 2017). Il prend également en compte une recherche ergonomique, c'est-à-dire une étude des gestes habituels, mais ce n'est pas systématique dans la mesure où le travail des ergonomes n'est possible que s'il existe déjà un produit, ou des produits similaires, dont on peut analyser empiriquement les usages. Il n'y a travail d'ergonomie que s'il existe des éléments mesurables (de confort, de lisibilité, etc.) au sein desquels il existe une marge de manœuvre. En l'absence d'un produit déjà existant (par exemple dans le contexte de la conception d'un produit présentant une innovation de rupture), le designer ne peut pas analyser les habitudes de ses utilisateurs, et doit se projeter dans des usages possibles qui procèdent essentiellement de son imagination et de son analyse de la problématique.

Dans le contexte du design industriel, il s'agit notamment d'analyser les autres modèles existants, les usages, les contextes dans lesquels l'objet est utilisé, l'environnement technique et économique, avant de proposer, en le décrivant ou en le fabriquant, un nouvel objet. Pour prendre un exemple relativement récent, les aspirateurs Dyson (**Figure 20**), récompensés par plusieurs prix de design, mobilisent un certain nombre d'innovations technologiques (la séparation cyclonique, l'absence de sac, un filtre à particules) tout en ayant une apparence singulière qui combine une recherche esthétique et une efficacité fonctionnelle. La forme d'un aspirateur Dyson se décrit difficilement par comparaison à d'autres aspirateurs, mais se comprend à partir de son schème de fonctionnement, qu'il matérialise formellement.

***Figure 20**. Photo d'un aspirateur Dyson (modèle Dyson DC52 Allergy Pro)*

Dans le contexte du numérique, le design est d'abord associé à ce qui est visible à l'écran : l'interface graphique. Il est ainsi assimilé à un travail de création (ou de modification) graphique qui consiste à rendre esthétique une interface déjà existante, créée par les développeurs. Le recours à un designer consiste bien souvent, en développement logiciel, à commander une maquette graphique une fois que le logiciel ou la fonctionnalité est développé(e). Cette tendance est largement dénoncée par les designers qui refusent d'être considérés comme des décorateurs d'interface, et revendiquent une place, voire un rôle central, dans le travail de conception. Stéphane Vial, maître de conférences en design, écrit ainsi :

> il faut dépasser le point de vue superficiel et tenace qui consisterait à croire que le design ne ferait qu'apporter un enjolivement visuel/formel ou une amélioration fonctionnelle qui apporterait simplement « un plus » au projet. (Vial 2016)

Dans le cadre du design d'interface, il est vrai que le travail du designer consiste à développer la lisibilité, l'utilisabilité d'une interface graphique et de faire en sorte qu'elle soit esthétiquement plaisante, de manière à rendre son utilisation agréable. Néanmoins, il ne se réduit pas à cela. Le rôle des interfaces hommes-machines, en général, est de permettre « la communication entre la composante humaine et la composante informatique du système » (Thierry 2013) ; une interface graphique est une interface qui remplit cette fonction avec des éléments visuels. Si l'esthétique, la lisibilité, l'utilisabilité, sont des critères qui favorisent la communication, la question centrale reste le contenu de cette communication, qui est déterminé en amont de son affichage à l'écran.

Nous rejoignons donc le point de vue soutenu par les designers et le refus de restreindre le design à une problématique d'ajustement qui interviendrait en fin de processus. Plus spécifiquement, nous considérons le design dans notre contexte comme une fonction non pas incarnée par une seule

personne, mais distribuée dans une équipe de développement logiciel. Il dépasse ainsi largement la fonction d'enjolivement qui lui est prêtée. Cette fonction consiste en effet à penser l'interface graphique, mais aussi ce qu'elle matérialise : c'est un travail de conception de l'invisible aussi bien que du visible, qui ne se réduit pas au design d'interface. Dans le contexte de la configuration en système et des moteurs de recommandation, le design est l'incarnation par excellence des préoccupations liées à ce que nous avons appelé l'écologie du projet. Il s'agit d'analyser les attentes des utilisateurs, leurs réactions à ce qui leur est recommandé, mais aussi de créer, parfois *ex nihilo*, de nouvelles formes de recommandation. Cette création relève de l'ingénieur, voire du data scientist, aussi bien que du designer, car il s'agit tout à la fois de concevoir techniquement un moteur de recommandation, mais aussi d'intégrer les usages et l'ambition de la recommandation. Dans cette conception, l'articulation avec les aspects techniques ne se fait pas seulement sous l'angle de l'intégration des contraintes techniques mais, au moins idéalement, dans un dialogue entre la proposition du design et les possibilités techniques : il peut certes s'agir d'intégrer telles quelles ces contraintes, mais aussi de renvoyer l'objet du côté de la recherche technologique, pour trouver des solutions techniques à une contrainte cette fois posée par l'ambition du produit.

À ce titre, le design désigne à la fois le travail de conception de la recommandation elle-même (design du produit) et de la façon dont elle va être donnée à voir pour l'utilisateur (design d'interface, d'expérience et d'interaction). La recommandation et sa représentation entretiennent une relation de codétermination où chacune influe sur l'autre. En effet, il faut que l'interface reflète ce que produit le moteur de recommandation, et mette en évidence les différentes fonctionnalités du logiciel. Néanmoins, à l'inverse, l'interface peut être déterminante dans la définition ou l'évolution de ces fonctionnalités dans la mesure où une fonctionnalité qui est difficilement matérialisable ou mal comprise par l'utilisateur devra en principe être transformée. Les fonctionnalités du logiciel ne sont pas seulement déterminées par ce qui est jugé pertinent ou efficace par les équipes de développement logiciel, mais aussi par le regard du design, l'analyse des comportements utilisateurs, et notamment leur réaction aux fonctionnalités matérialisées par l'interface.

De ce fait, il reste vrai qu'une interface graphique peut s'évaluer sur des critères d'esthétique ou de lisibilité, mais elle doit être avant tout ce qui donne accès aux fonctionnalités et rend possible leur utilisation, ou autrement dit, qui « donne à voir » et « permet d'agir », pour reprendre le titre de la thèse de Benjamin Thierry (2013) sur l'invention de l'interactivité graphique et du concept d'utilisateur en informatique. Pour cela, toute interface graphique s'articule autour de deux concepts clés : l'affichage, qui matérialise des éléments du système, et l'interactivité, qui permet d'agir sur ces éléments. Utiliser un logiciel, c'est successivement consulter des éléments d'affichage et manipuler des éléments interactifs suivant un ordre ou une méthode déterminés par une intention. Dans le cadre des systèmes de recommandation, les deux éléments essentiels sont donc l'affichage des suggestions (de produits, de contenus, de films, etc.) et la possibilité de choisir une (ou plusieurs) d'entre elles,

généralement en cliquant dessus. À ce titre, une interface comme celle de Netflix[55] (**Figure 21**) combine de multiples zones de recommandations que nous détaillons ici :

- une grande mise en avant dynamique (personnalisée en fonction de l'utilisateur, du pays, des entrées au catalogue, de la stratégie éditoriale de Netflix, etc.) d'une vidéo en particulier, qui peut être un film, une série, un documentaire, etc. ;
- une liste de listes de recommandations, suggérant de reprendre la lecture d'une vidéo ou de parcourir plusieurs listes de suggestions ;
- une fonctionnalité de recherche (tronquée sur la **Figure 21**) qui va en pratique suggérer des vidéos en rapport avec les termes de la recherche, soit des titres contenant les termes ou une partie des termes, soit des termes similaires, etc., c'est-à-dire, d'autres suggestions de vidéos ;
- une zone de notification (également tronquée sur la **Figure 21**) des nouvelles entrées au catalogue, mais sous formes de suggestions personnalisées pour l'utilisateur.

***Figure 21**. Captures d'écran de deux variantes de la page d'accueil de Netflix présentées dans The Netflix Recommender System : Algorithms, Business Value, and Innovation (2015) un article de Carlos Gomez-Uribe et Neil Hunt, tous deux travaillant chez Netflix. La première mise en avant varie (The Others dans la première capture d'écran, The Universe: Collection dans la seconde). La première liste de recommandations est d'un côté une liste de « films à suspense », de l'autre les « top picks » pour l'utilisateur. Les listes de recommandations suivantes sont également différentes.*

En dehors des fonctionnalités liées à la gestion du compte, et quelques éléments de navigation statique, l'interface est presque essentiellement constituée de zones d'affichage de recommandation, hiérarchisées par l'espace qu'elles occupent à l'écran et par l'ordre dans lequel elles apparaissent. Cet ordre d'affichage reflète quant à lui l'ordre calculé par le moteur de recommandation, sous la forme d'un score invisible à l'écran. L'élément d'interactivité est quant à lui presque toujours le même : un bouton « play » soit déjà visible, soit apparaissant lorsqu'on clique sur une vidéo donnée. Affichage et interactivité sont centrés sur la consultation de recommandations et l'action d'en choisir une.

Comme tant d'autres interfaces, celle de Netflix matérialise ainsi deux types d'éléments, éventuellement superposés : des éléments d'affichage, qui permettent à l'utilisateur de voir et de contrôler, et des éléments d'interactivité, qui lui permettent de déclencher une action. Voir et agir

[55] Nous présentons ici l'interface web. L'application pour tablette et smartphone présente des variations que nous n'allons pas détailler ici.

sont les deux facettes de Netflix en tant que produit, c'est-à-dire en tant que somme des usages rendus possibles par l'application. La définition et la conception de ce qu'est ce produit émerge conjointement du développement et de l'évolution de ses fonctionnalités techniques, et de ce qui est matérialisé à l'utilisateur par l'interface. Le produit n'est pas l'un ou l'autre mais ce qui émerge de la somme et des interactions entre les deux.

On retrouve ici le statut de l'informatique comme transcendantal empirique délimitant un espace de possible. Au chapitre V, nous avions vu en effet que l'informatique délimitait pour le data scientist l'espace de la calculabilité, c'est-à-dire un espace de possible computationnel. De la même façon, au chapitre VII, le logiciel programmait un ensemble d'actions possibles et répétables qui constituent l'espace au sein duquel l'utilisateur peut évoluer. Plus précisément, nous pouvons maintenant dire que c'est le produit composé des fonctionnalités du logiciel (elles-mêmes basées sur un travail de *data science*), et de son interface, qui détermine les usages possibles (sans jamais les anticiper tous absolument), et matérialise leur résultat. Les éléments d'interface permettant l'interactivité sont comme des suggestions d'action possible. Ils déterminent ce que l'utilisateur peut faire, mais jamais ce qu'il va effectivement faire. C'est un fait bien connu en effet de tous ceux qui incarnent la figure du designer, que les usages d'un produit existant débordent presque systématiquement du modèle d'usage anticipé, et que les produits sont détournés, recontextualisés, affectés à d'autres emplois que ceux prévus. Parce que les usages effectifs échappent à l'anticipation, le suivi *(tracking)* des actions utilisateur est une tâche nécessaire pour comprendre ces usages, leur décalage avec les usages anticipés, les typologies d'usages et d'utilisateurs, et ainsi faire évoluer le produit.

Celui-ci ne fait toujours que suggérer des usages, sans les imposer ou les programmer, suivant quoi on peut désormais généraliser la notion de logiciel de recommandation abordée dans la configuration en système à toute interface logicielle. Recommander, c'est proposer différents choix (éventuellement hiérarchisés) qui constituent déjà une sélection par rapport à l'ensemble des choix possibles : si j'indique à un ou une amie des films à voir, que je lui recommande, je ne liste pas la totalité des films que j'ai vus. En élargissant la notion de recommandation, non pas à la seule recommandation de contenus fondées sur l'analyse computationnelle de données numériques, mais également à la recommandation en général, un produit logiciel est ce qui recommande et rend possible des actions, parmi une infinité d'actions qui auraient pu être développées comme fonctionnalités et matérialisées. En ce sens, tout produit logiciel est un système de recommandation d'action, quand bien même la seule proposition d'action serait d'activer le produit.

La différence entre les produits de recommandation fondés sur les sciences des données, tels que nous les avons abordés en présentant la configuration en système, et les produits logiciels en général, est leur **caractère génératif**. Dans un logiciel classique en effet, les éléments d'interactivité sont majoritairement statiques. Il est relativement facile d'anticiper l'état de l'interface à un moment donné pourvu qu'on dispose des informations associées à l'utilisateur (ses paramètres, et l'historique de ses

actions). Dans le cas d'un moteur de recommandation au contraire, il est difficile de savoir ce que propose l'interface si ce n'est en la consultant effectivement. La seule façon de reconstituer ce qui sera affiché est de rejouer le calcul qui produit la recommandation. Il faut pouvoir manipuler le moteur, et disposer de toutes les informations qui sont prises en compte dans le calcul. Il n'est pas possible d'anticiper le résultat : le calcul effectif est nécessaire. Même un développeur logiciel travaillant sur le produit, ayant accès au code et aux serveurs de production, ne pourra pas déterminer le résultat du calcul autrement qu'en reconstituant le processus de génération de la recommandation. Or la manière la plus simple de reconstituer cette génération est de consulter l'interface, qui est virtuellement singulière par la somme des paramètres suivant lesquels elle est générée. De ce fait, si notre présentation de la configuration en système présentait de nombreuses similarités avec l'activité de développement logiciel en général, les produits logiciels fondés sur les sciences des données s'en distinguent en ce qu'elles intègrent une phase de recherche et développement (l'étape de concrétisation interne simondonienne qui aboutit à la mise au point du moteur de recommandation) dont le résultat matérialisé dans l'interface produit sous forme de suggestions des résultats conçus d'une telle manière qu'ils ne peuvent pas être anticipés. Les interfaces des produits logiciels fondés sur les sciences des données matérialisent des informations qui ne peuvent pas être anticipées, et ne peuvent être déterminées qu'en rejouant le processus qui les produit (ou des processus équivalents).

Par sa capacité à donner à voir ces informations, l'interface joue dans les produits logiciels de recommandation un rôle similaire à celui de la visualisation de données dans la configuration en récit. En effet, comme la visualisation, elle matérialise des éléments de sens mais ne porte pas en elle-même une interprétation de leur signification : c'est le système composé de l'interface et de l'utilisateur qui est capable de produire une telle signification. Les interfaces de produits logiciels grand public sont généralement proposées sans explication ni manuel d'utilisation ; l'utilisateur apprend seul à l'utiliser et doit comprendre par lui-même ce qui lui est présenté. La présence d'une explication aux recommandations elles-mêmes est encore loin d'être systématique, et cette explication est généralement assez elliptique. Les publications de mes contacts Facebook, mes résultats de recherche sur Google, me sont donnés tels quels. À l'inverse, Amazon m'indique que les produits recommandés le sont car d'autres personnes consultant la même page que moi ont consulté ou acheté ces produits ; Netflix précise qu'il me recommande soit un certain genre de vidéos (des thrillers, des films d'auteur, etc.), soit des sorties récentes, soit d'autres types de recommandations. Les spécialistes de la recommandation avec lesquels nous nous sommes entretenu s'accordent à dire que proposer une explication des recommandations améliore significativement le déclenchement d'une action de l'utilisateur : la recommandation en est plus lisible, plus facile à comprendre. Ces explications jouent donc un rôle similaires aux éléments textuels (libellés, titres, annotations, etc.) ajoutés aux visualisations de données en vue de leur restitution. Comme la visualisation, l'interface qui présente des suggestions (et des explications éventuelles) constitue une forme visuelle générée

computationnellement, présentant des éléments sémiotiques conventionnels, mais aussi un supplément de sens non explicité qui n'apparaît que par l'interprétation, et n'existe donc qu'en anticipation d'un utilisateur.

5. Matérialisation et interprétation de la recommandation

Les recommandations générées par les produits logiciels le sont sur un temps qui est distinct du temps de conception et de développement (et occasionnellement, du temps de consultation) ; contrairement à la configuration en récit, il n'y a pas un analyste qui défend et justifie en personne les résultats de son travail. Comme on l'a noté, des explications aux éléments proposés sont parfois disponibles, mais ces explications elles-mêmes sont générées par un calcul : virtuellement, un analyste qui voudrait justifier une liste de recommandations devrait également justifier l'explication générée avec celles-ci. De ce fait, les explications comme celles que proposent Netflix laissent une large place à l'interprétation et à d'éventuelles justifications, voire, comme nous allons l'esquisser, à une remise en cause des explications et des recommandations qu'elles accompagnent.

Ainsi, lorsque Netflix m'indique par exemple que l'une des listes de films qu'il me recommande est une liste de films historiques, il ne m'explique pas pourquoi il me recommande ce genre, ni comment ces films ont été catégorisés comme tels. Sur un mode abductif, l'utilisateur doit formuler des hypothèses explicatives pour comprendre ce choix de recommandation : je peux supposer qu'il me suggère des films historiques car le dernier film que j'ai vu en était un, parce que j'en ai vu plusieurs dans les 30 derniers jours, ou encore parce que c'est le genre de films que j'ai le plus regardé depuis que j'ai un compte. En l'absence d'une justification à cette catégorisation, je peux imaginer toutes sortes d'explications. Par ailleurs, la classification en film historique des films recommandés ne m'est pas expliquée non plus : a-t-elle été faite manuellement par un spécialiste de cinéma ou du genre ? S'agit-il d'un consensus de plusieurs spécialistes ? S'agit-il d'une catégorie fournie par le réalisateur, par les salles de cinéma ? Est-ce le libellé ajouté manuellement à un regroupement automatique de films qui présentaient des similarités, généré par un algorithme ? Quels étaient les paramètres de ce regroupement ? Si la catégorie « film historique » peut apparaître comme une catégorie naturelle et évidente, l'irruption de cas limites ou de désaccords remet en cause une posture initialement naturaliste.

Ainsi, sur la **Figure 22**, le premier film d'une liste de recommandations libellées « films historiques » est le documentaire diffusé en mai 2017 sur la campagne présidentielle d'Emmanuel Macron. En termes de genre audiovisuel, de période historique visée, de régime d'historicité, le fait de mélanger documentaire et film historique est largement sujet à discussion : on peut aisément contester qu'il s'agisse de catégories identiques, ou appréciées par les mêmes personnes, mais on peut également leur trouver des caractéristiques similaires. Dans tous les cas, l'apparition de ce cas limite dénaturalise une catégorie qui aurait pu apparaître homogène sans cet élément.

Figure 22. *Capture d'écran de l'interface de Netflix, montrant une partie de la liste de « films historiques » recommandés. La première proposition n'est pas un film historique mais un documentaire (*Emmanuel Macron, les coulisses d'une victoire*) portant de plus sur un sujet contemporain.*

À travers un exemple de dysfonctionnement, on voit comment la signification des recommandations n'est pas donnée par l'interface, mais construite par l'interprétation de l'utilisateur. Les regroupements proposés ne correspondent pas tant à des genres institutionnalisés qu'à des catégories émergentes auxquels il faut trouver un sens, et qui sont autant de clés d'analyse en puissance des contenus audiovisuels. À un niveau méta, ces recommandations constituent également des suggestions de typologie de ces contenus. Au contraire des catégories comme le genre cinématographique, qui sont des catégories explicitement définies et conceptualisées par des experts, avec des critères d'identification plus ou moins fins, institutionnalisées au point que le mélange des genres cinématographiques et le jeu de références à différents genres est devenu un phénomène courant de la création audiovisuelle, les catégories émergentes de la recommandation ne sont définies que par l'exemple. En ajoutant à ces listes de recommandation une indication de genre cinématographique, Netflix tente de rattacher ses catégories émergentes, constituées par l'exemple, à des catégories existantes, ou en d'autres termes, de ramener du particulier à du général : l'interface propose non seulement des recommandations de contenus audiovisuels, mais des suggestions de rattachement de ces contenus à des catégories plus générales. Elle rejoue et matérialise ainsi le problème classique de la classification supervisée, qui s'efforce de relier par le calcul le divers de la donnée (plutôt que le divers de l'expérience) à des concepts connus.

En considérant le concept de genre cinématographique comme une forme de savoir sur l'objet, nous proposons de décrire ces catégories émergentes comme des **connaissances tacites et infra-conceptuelles** qui peuvent devenir des connaissances effectives par un travail d'interprétation de l'utilisateur. Au lieu d'un raisonnement abductif, à travers lequel l'utilisateur formule une hypothèse explicative pour les recommandations qui lui sont faites, dans le but de comprendre les éléments de l'interface avant de faire son choix, on peut en effet appliquer un raisonnement inductif à ces catégories émergentes pour les conceptualiser et les amener à un niveau de généralité supérieur. Ce travail nécessite naturellement une expertise qui n'est pas simplement celle de l'utilisateur, mais d'un spécialiste du domaine qui va pouvoir mobiliser des connaissances extérieures et des schémas conceptuels ou narratifs. Dans cette situation virtuelle, qui détourne l'usage prévu des systèmes de recommandation, mais qui permettrait de réintégrer l'épistémologie des sciences de la culture dans

les configurations en système, on est en réalité proche du travail de l'analyste dans la configuration en récit : à partir d'éléments matérialisés visuellement, l'interprète élabore des significations possibles qui acquièrent une systématicité par un travail de justification et d'auto-falsification des catégories proposées, restitué non plus sous la forme d'une interface, mais sous forme de récit.

Il y a donc de multiples circulations possibles entre la configuration en récit et la configuration en système, y compris dans leurs modalités respectives de visualisation de données. On peut d'ailleurs voir l'interface comme un récit spatialisé, qui réaliserait une synthèse non pas temporelle, mais spatiale, de l'hétérogène, selon les termes de Ricœur, et laisserait du jeu dans la mécanique narrative pour inclure le lecteur dans sa construction. Quelle que soit la configuration, la visualisation de données joue le rôle d'une sémiotique à la fois heuristique, permettant de produire des connaissances, et rhétorique, facilitant leur explicitation et leur explication. Par son caractère initialement génératif, elle joue également un rôle de preuve par fondation empirique dans la mesure où elle donne à voir les données, c'est-à-dire des éléments de représentation du réel. Elle préserve et révèle le sens des données, construit par le calcul et les choix de l'analyste, mettant en évidence une interprétation dont la validité procède du destinataire de l'analyse, et de son contexte épistémique.

Chapitre IX.
Connaissance et décision

Tout au long des chapitres précédents, nous avons vu comment la connaissance émergeait, quelle que soit la configuration, d'une combinaison d'actions de préservation et de révélation à partir de données auxquelles on attribue une valeur épistémique. Les transformations de la donnée passent par un grand nombre d'étapes non linguistiques et infra-conceptuelles, qui relèvent du calcul puis de la visualisation, et que la mise en langage du résultat de ces transformations n'intervient qu'à la fin du processus, pour en tirer le sens. D'un côté, les traitements computationnels *préservent* cette valeur, tandis que les choix de l'analyste, et les interprétations qu'il formule, *révèlent* le sens de la donnée. Ainsi, ces activités fondent les connaissances tirées des données de deux façons : elles les légitiment, en apportant un élément de preuve, mais surtout, elles donnent un statut d'assertion à quelque chose d'auparavant inintelligible. Dans les termes de la philosophie analytique, on dira qu'elles ont à la fois un rôle de vériporteur et de vérifacteur (Armstrong 1993).

En effet, pour que les connaissances tirées des données soient valides, il faut d'abord qu'elles soient des connaissances, c'est-à-dire des assertions qui peuvent être falsifiées et auxquelles on accorde un degré de croyance. La visualisation est un support de la connaissance ou vériporteur *(truth-bearer)* qui pose les termes dans lesquels nous formulons des connaissances valides ou invalides ; l'interprétation est une assertion falsifiable formulée à partir de cette visualisation, c'est-à-dire une connaissance hypothétique. À ce titre, la visualisation joue un rôle similaire à celui du modèle en sciences :

> According to the semantic view, scientific knowledge is embodied in models, which make statements true, but are not themselves true or false. (Perini 2012)

En d'autres termes, la visualisation est la structure dans laquelle les connaissances sont matérialisées ; c'est une assertion qui peut être vraie ou fausse. Plus précisément, c'est le jugement porté sur la visualisation (c'est-à-dire son interprétation) qui peut être validé ou falsifié et constitue une assertion, tandis que la visualisation pose les termes dans lesquels cette assertion est posée.

Néanmoins, la visualisation est également une source de connaissance ou un vérifacteur *(truth-maker)* dans la mesure où, comme on l'a vu précédemment, elle joue un rôle de preuve de ce qu'elle représente. Il ne s'agit donc pas seulement de formuler une connaissance hypothétique, dont la valeur de vérité est inconnue, mais de justifier cette connaissance et d'appuyer cette valeur de vérité au moyen de la visualisation de données.

Toutefois, l'emprunt que nous faisons à la philosophie analytique avec les notions de vérifacteur et vériporteur se place dans une conception plutôt réaliste où les connaissances tirent leur régime de vérité dans leur correspondance avec le réel. Or cette posture réaliste s'accorde mal avec le statut que nous avons donné, à travers les différents chapitres qui précèdent, à la donnée et ses traitements. La valeur épistémique de la donnée est construite dans son geste constitutif de sélection ; les traitements s'évaluent sous l'angle de l'objectivité comme préservation et révélation de cette valeur. Cette valeur épistémique n'était jamais posée comme correspondance au réel, celle-ci ne pouvant jamais être autre chose que supposée du point de vue de l'analyste. C'est précisément le propre de la posture de l'analyste que de faire des données ce qui est (ou ce qui est donné), le monde à explorer indépendamment de son éventuel correspondance au monde réel. La donnée peut occasionnellement être posée comme une correspondance au réel, mais elle est surtout toujours un indice de celui-ci, qui doit être manipulé et interprété pour être reconnectée à l'événement qui l'a générée, et faire sens.

De ce fait, la posture réaliste qui était par défaut celle de la mythologie des *big data* ne peut être soutenue dès lors que l'on prend en compte leurs modalités et leur régime de constitution de connaissances. Au regard des remarques déjà formulées sur le statut des connaissances produites, nous proposons de les évaluer au prisme d'une approche non pas réaliste, mais plutôt d'ordre pragmatiste, qui articule de manière resserrée les notions de connaissance et d'action.

1. Les *big data*, une épistémologie pragmatiste ?

Si l'on prend en compte leur contexte d'utilisation, la finalité des data sciences n'est pas en effet de dire le vrai. Si les résultats des traitements computationnels peuvent occasionnellement être vus comme porteurs de vérité, ce n'est pas dans ce but qu'ils sont réalisés : la production même de connaissances y est instrumentale, au service d'autre chose. De ce fait, elles peuvent être analysées et interprétées dans un cadre réaliste, mais également, comme nous allons maintenant le voir, dans un cadre instrumentaliste. Si les deux interprétations coexistent, et apportent chacune une compréhension de la nature des connaissances produites, nous allons voir que l'interprétation instrumentaliste régit l'interprétation réaliste. Bien que la véridicité elle-même ait une fonction dans la validation des connaissances et notamment dans la confiance qui leur est portée, c'est le caractère instrumental de la connaissance issue des données en général (et non pas d'une connaissance instrumentalisée pour en produire une autre, comme on le trouve dans les épistémologies instrumentalistes) qui caractérise les sciences des données par rapport à d'autres pratiques

scientifiques. Ancrées dans le contexte d'une entreprise, elles s'articulent en effet avec la finalité de celle-ci : sa survie et sa croissance. Les connaissances produites par une entreprise ne le sont, en principe, qu'en vue de l'une de ces finalités. Cette articulation entre connaissance et finalité de l'entreprise se fait par la notion de décision, et plus précisément par la capacité des connaissances issues du traitement de données massives à améliorer la prise de décision. La finalité de ces traitements serait ainsi de prendre des décisions régies par des données *(data-driven decisions)*. Dans un article qui tente de clarifier les liens entre sciences des données et décision, Foster Provost et Tom Fawcett (2013) écrivent ainsi :

> Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data. For the perspective of this article, the ultimate goal of data science is improving decision making, as this generally is of paramount interest to business. […] Data-driven decision making (DDD) refers to the practice of basing decisions on the analysis of data rather than purely on intuition.

La promesse des sciences des données n'est ici pas tant de produire des connaissances nouvelles, que de permettre aux entreprises d'agir de manière plus rationnelle, en fondant leurs décisions empiriquement et non sur la seule intuition. Plus bas, ils soulignent que cet empirisme a des conséquences pratiques, qui justifient le recours à une telle approche :

> The benefits of data-driven decision making have been demonstrated conclusively. Economist Erik Brynjolfsson and his colleagues from MIT and Penn's Wharton School recently conducted a study of how DDD affects firm performance. They developed a measure of DDD that rates firms as to how strongly they use data to make decisions across the company. They show statistically that the more data-driven a firm is, the more productive it is—even controlling for a wide range of possible confounding factors.

En ces termes, l'empirisme fondé sur les données est motivé par les conséquences pratiques qu'il a sur la survie et la croissance de l'entreprise, à travers l'amélioration de sa performance : des données sont analysées dans la perspective d'œuvrer pour cette finalité. La valeur dans l'absolu des résultats produits est accessoire, ou du moins n'a d'intérêt qu'en tant que ces résultats suscitent une décision bénéfique pour l'entreprise. C'est un antiréalisme, non pas au sens où ce point de vue rejette une interprétation réaliste de ces résultats, mais au sens où il est indifférent à son caractère réaliste.

On pourrait aller jusqu'à dire que la validité des résultats est déterminée non pas sur des critères scientifiques, mais précisément du point de vue des bénéfices que ces résultats peuvent apporter. L'épistémologie des sciences des données s'apparenterait alors à une forme de pragmatisme, c'est-à-dire une théorie de la connaissance fondée sur la maxime pragmatiste de Peirce (1878) :

> Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object.

Les commentateurs de Peirce (Tiercelin 1993; Hookway 2013) ont souligné combien Peirce explicite peu la démarche pour déterminer ces conséquences pratiques. Néanmoins, on peut compléter cette citation en explicitant l'état d'esprit général du pragmatisme peircéen, qui consiste à rejeter les énoncés métaphysiques et à envisager les connaissances du point de vue de ce qu'elles signifient concrètement, par rapport à des problèmes qui se présentent effectivement. En ces termes, une connaissance est validée non pas à partir d'un cadre métaphysique, mais au regard d'une situation qu'elle permet ou non de résoudre. Nous comprenons ainsi que non seulement la connaissance est ce qui permet l'action, mais une connaissance n'est connaissance que parce qu'elle rend possible une action.

En ces termes, les sciences des données s'apparentent à une forme spécifique de pragmatisme, où les connaissances issues des traitements de données massives ne sont connaissances que parce qu'elles facilitent une prise de décision dans une entreprise et que cette prise de décision va d'une façon ou d'une autre se traduire en action. Elles apparaissent comme une forme plus restreinte de pragmatisme, articulée à la notion de prise de décision et non à celle d'action en général. En effet, une prise de décision est une forme d'action, mais la production de connaissances en est une aussi : la notion d'action est plus large que la notion de prise de décision, qui rend l'épistémologie des sciences des données plus restrictive que le pragmatisme dans sa conception des connaissances. Elle ne serait pas seulement un pragmatisme au sens *ordinaire*, c'est-à-dire une forme de « bon sens » qui privilégierait ce qui peut être mis en œuvre facilement au détriment d'un travail théorique et philosophique, ni un pragmatisme *linguistique* comme on l'a vu à l'œuvre dans la codification de verbatims, mais une théorie de la connaissance régie par le concept d'action, apparentée au pragmatisme *philosophique*, d'autant plus fortement articulé au concept d'action qu'elle serait une philosophie pratique opérationnalisée et pas seulement conceptualisée.

Néanmoins, l'hypothèse d'une interprétation pragmatiste demande à être discutée. Elle suppose à première vue qu'il existe un lien quasi-mécanique entre connaissance et action. Or, si le concept de décision qui les articule est comme on va le voir, central dans les processus de traitement de données, il s'entend dans un sens bien précis dont il n'est pas certain qu'il propose une articulation mécanique, et de ce fait, que les connaissances produites relèvent effectivement du pragmatisme au sens où Peirce l'entend.

Il nous faut donc commencer par examiner précisément le concept de décision, en général et dans le contexte des *big data*. Dans son sens général, une décision est le résultat d'un processus qui implique l'évaluation d'un ou plusieurs choix possibles (la délibération), et la sélection d'un choix. La délibération est une action discursive, individuelle ou collective, qui permet de poser un jugement. La délibération et la sélection sont des processus mentaux qui n'ont pas de conséquence mécanique : décider d'acheter un appartement n'aura pas pour conséquence de me rendre propriétaire. Ces processus mentaux reflètent également l'état de mes croyances à un moment donné : je vais opter

pour l'une des possibilités car elle me semblera plus désirable au vu de mes connaissances et de mes critères moraux à ce moment-là. Une décision peut être vue comme une action (de l'esprit) qui va entraîner d'autres actions. Néanmoins, le rapport entre une décision et une action peut être lointain : je peux tout à fait décider quelque chose et ne jamais mettre à exécution ma décision. Lorsqu'un dirigeant d'entreprise présente une stratégie, il énonce un ensemble de décisions qui devront être reformulées et décomposées à de multiples reprises avant d'être mises en œuvre. À posteriori, on peut évaluer les actions associées à cette mise en œuvre et leur adéquation à la décision initiale ; de cette façon il peut arriver qu'on croie mettre en œuvre une décision, mais qu'en réalité la traduction de cette décision en action soit erronée. Dans cette situation, il y a bien une action qui effectuée, mais qui ne correspond pas à la mise en œuvre de la décision visée. Aussi, indépendamment du fait que décider est une activité en soi, une opération mentale, le fait qu'une action soit la conséquence pratique d'une décision n'a rien d'évident, et peut faire par exemple l'objet d'une évaluation *a posteriori* pour s'en assurer.

Or, ce concept de décision qui sert à articuler connaissance et action est une notion clé des sciences des données qui, comme on l'a vu, se veulent initiatrices de prises de décision fondées sur les données. Dans sa polysémie, il traverse l'ensemble des processus de traitements de données massives. Ainsi, si l'on retrace ces différentes étapes :

- le **geste de constitution des données** est une décision, d'apparence arbitraire ;
- l'**informatique théorique**, et tout calcul par la suite, répond au problème de la décision de Hilbert, qui consiste à poser un problème sous la forme d'une alternative entre plusieurs choix parmi lesquels il faut trancher. L'informatique théorique est alors l'art de concevoir une procédure effective qui permet de trancher entre ces différentes options, et ainsi de résoudre le problème en faisant de la délibération une procédure computationnelle ;
- les analystes et développeurs ne cessent de prendre des décisions, de faire des **choix de conception** quant à la bonne façon de nettoyer, analyser et visualiser les données ;
- l'utilisateur de la configuration en système décide d'un choix parmi les **recommandations** qui lui sont faites ;
- comme on vient de le voir, les résultats de ces traitements ont pour fonction d'amener à de meilleures **prises de décision**.

À travers ces différents points se dégagent des conceptions différentes de la décision, que nous allons détailler avant d'approfondir celle qui pourrait matérialiser la connexion entre connaissance et action.

- Tout d'abord, les décisions prises par la détermination du calcul sont prises par des agents non-humains : le fait de dire qu'un algorithme, et le programme qui l'implémente, « décident » des contenus qui me seront proposés *via* un système de recommandation peut être vu comme un abus de langage. Il convient *a minima* de souligner que les décisions

algorithmiques et les prises de décision humaines ont un statut ontologique distinct. Les décisions des programmes ne sont pas celles qui permettent à l'entreprise d'accomplir sa finalité, mais des déterminations intermédiaires qui permettent à un utilisateur spécifique, le décideur, de se voir proposer plusieurs options. L'algorithme joue en effet le rôle d'une procédure mécanique en vue de la résolution d'un problème *posé comme* un problème de décision : cette procédure permet de décider par le calcul des choix qui seront soumis à l'utilisateur. De ce fait, on peut voir les systèmes de recommandation comme des systèmes d'*aide à la décision* (et non des systèmes de décision) qui réalisent une présélection parmi l'ensemble des choix possibles. La logique et les paramètres de cette présélection sont déterminés par un calcul lui-même issu d'une multitude de choix de conception et d'implémentation de la part de l'analyste ou du développeur. Cette présélection est un support à la phase de délibération qui précède la sélection d'un choix mais le calcul lui-même n'est pas une délibération. On parlera pour cette étape de *déterminations* plutôt que de *décisions* computationnelles.

– Les décisions épistémiques de l'analyste ou du développeur sont quant à elles des choix d'apparence arbitraire dans du non-nécessaire, régis par une logique (d'acquisition, de transformation), qui ont des conséquences directes sur les connaissances produites. Ce sont bien des décisions au sens premier que nous avons proposé : des opérations mentales qui aboutissent à la sélection d'un choix parmi plusieurs options possibles. Ces choix sont effectués en vertu d'une capacité à produire des jugements et manipulations exercées préservant ou révélant les vertus épistémiques de la donnée. Sans ces choix, la donnée demeure inintelligible : la prise de décision est une étape clé de la production de connaissances par le traitement de données massives.
– Au bout de la chaîne de transformation, ces connaissances sont mobilisées par des utilisateurs afin de faciliter une décision. Dans la configuration en système, la recommandation propose différentes options tandis que le logiciel rend opérationnalisable la sélection de l'une de ces options. Néanmoins, ce schème de la recommandation, qui consiste à présélectionner des options, existe également dans la configuration en récit. Les consommateurs de récit s'attendent en effet à ce que le récit présente non seulement les conclusions de l'investigation, mais que découle de ces conclusions des recommandations d'action : le type de connaissance attendu n'est pas une synthèse qui rend compte des données, mais des propositions d'action informées par cette synthèse.

Ce rapport entre conclusion et recommandation est un point important dans la revendication d'une éventuelle épistémologie pragmatiste permettant d'articuler une relation entre connaissance et action. Les traitements de données s'appuient d'une part sur des décisions d'agents humains et non-humain, mais c'est cette étape de recommandation qui présente une articulation entre les connaissances

produites (et non la production de connaissances) et l'action médiée par la décision. Nous allons l'examiner plus en détail.

Comme on l'a vu précédemment, la configuration en récit peut prendre la forme d'un rapport d'étude remis et présenté à un client. C'était notamment le cas des services proposés par Proxem. Or, à plusieurs reprises, j'ai observé en produisant et en présentant des rapports de ce type, que les clients de Proxem ont pu être déçus par le rapport d'étude produit, jugé « inutile », peu « parlant ». Les conclusions n'étaient pas remises en cause (ce n'était donc pas leur validité épistémique qui était contestée) mais le client ne voyait pas en quoi ni comment il pouvait s'en servir. En termes pragmatistes, on pourrait dire que ces conclusions n'étaient pas des connaissances dans la mesure où elles ne permettaient pas au client d'en tirer des conséquences pratiques, des décisions d'action.

À la suite de ces réactions, les rapports produits et visés ont été remaniés de manière à reformuler la synthèse sous formes de propositions d'action, par exemple en transformant une conclusion comme

> « Les interruptions de service sont une source importante d'insatisfaction pour les clients. »

en la proposition

> [sous-entendu, « nous vous recommandons de »] « Proposer un geste commercial en cas d'interruption du service. »

Les rapports ainsi mis à jour devenaient un récit toujours fondé sur la visualisation de données, mais dont la conclusion reprenait les éléments narratifs produits tout au long du récit pour les reformuler en propositions d'action.

Soulignons ici que l'introduction d'éléments de recommandation dans un rapport d'analyse peut surprendre au regard des normes des publications scientifiques, mais reste tout à fait ordinaire dans le monde professionnel. Dans les métiers du conseil en effet, les propositions d'action sont appelées des *recommandations* (voire des *recos*) ; c'est un terme stabilisé pour décrire l'apport des sociétés de conseil, dont la fonction n'est pas tant d'apporter une expertise que de proposer des actions sur la base de cette expertise. Les attentes des clients de Proxem vis à vis des rapports d'études produits relèvent ainsi davantage de la société de conseil que de l'institut d'étude, dont le rôle est initialement de réaliser une enquête et de la synthétiser[56]. De ce fait, un rapport se terminant par des recommandations est systématiquement perçu comme plus utile.

Cette utilité, qui reflète bien une conception instrumentaliste de la connaissance où la validité de celle-ci ne repose pas (ou pas seulement) sur sa valeur de vérité, est souvent exprimée à travers un terme

[56] De plus en plus, les instituts d'études sont également soumis à l'injonction de produire des recommandations. Cette injonction s'accompagne d'une restructuration du marché par laquelle ces instituts d'études sont de plus en plus intégrés à des grands groupes d'agences conseil en communication. Les différentes entités du groupe sont ainsi en mesure de produire l'analyse mais aussi la recommandation, ainsi que de la mettre en œuvre de manière opérationnelle.

récurrent de l'univers des *data-driven decisions :* la notion d'**actionnabilité**. Issue de l'anglais *(actionability)*, cette notion qualifie ce qui peut donner lieu à une action, ce sur la base de quoi on peut agir. Une connaissance est actionnable dès lors qu'on peut facilement identifier une action qui découle de cette connaissance. *Activer* une connaissance actionnable, c'est la mettre en œuvre. Dans un cadre pragmatiste, une connaissable est par définition actionnable : elle n'est connaissance que si elle peut avoir des effets pratiques, des conséquences relatives à l'action. En exigeant que les récits s'achèvent sur des recommandations plutôt que des conclusions synthétiques, descriptives de l'intelligibilité des données, les destinataires de ces récits inscrivent les connaissances produites par les analystes dans un cadre pragmatiste. Les actions à produire étant celles de ces destinataires, eux seuls peuvent effectivement attester de leur actionnabilité : l'analyste peut imaginer, anticiper des possibilités d'action correspondant à ce qu'il a fait émerger des données, mais c'est le destinataire décisionnaire qui pourra déterminer si une connaissance donnée devra être suivie d'action.

Entre la recommandation et l'action, il y a donc bien un terme tiers qui les relie, et qui n'appartient pas au domaine d'action de l'analyste : c'est le domaine de la décision dans son sens 3). Comme d'autres étapes antérieures, il mobilise à son tour, comme on va le voir, une capacité d'interprétation et de jugement complétée par des connaissances externes.

2. Vers une validité graduelle des connaissances

Tout au long du processus de traitement de la donnée, tout se passe comme si les connaissances potentielles contenues dans les données étaient actualisées. L'interprétation, le calcul, actualisent des assertions qui ne sont pas inventées *ex nihilo* (elles proviennent de la donnée) mais qui ne sont pas non plus manifestes à travers la donnée « brute », c'est-à-dire la donnée qui n'a pas encore été manipulée. Les décisions prises par l'analyste ou le développeur, et les « décisions » computationnelles des programmes, sont les moteurs de ce processus d'actualisation des connaissances. Néanmoins, ce processus d'actualisation est incomplet sans une étape de validation, qui rend les connaissances effectives, les valide comme connaissances. Or, dans un cadre pragmatiste, ce qui valide ces connaissances est le fait qu'il est possible de les mettre en œuvre, d'agir sur la base de ces connaissances. L'agent épistémique en mesure de déterminer si une action est possible est le destinataire du récit (ou l'utilisateur du système) : sans lui, la connaissance n'est toujours qu'en puissance, car c'est sa capacité à agir qui est en jeu pour rendre effectives les connaissances.

Suivant cette logique, on peut alors se demander si une connaissance pragmatiste est connaissance a) dès lors qu'on admet qu'elle *peut* donner lieu à une action, b) dès lors qu'elle peut donner lieu à une *décision d'action*, ou encore c) dès lors l'action a effectivement été *mise en œuvre*. Dans une configuration en système, ces trois possibilités peuvent être très rapprochées, voire indistinguables : le choix est toujours déjà possible et opérationnalisable, programmé par l'interface, accessible en un clic. La seule part qui n'est pas fournie par le système est l'atomicité essentielle de la décision de

l'utilisateur, ainsi que la délibération éventuelle qui peut la précéder. Cette décision n'est d'ailleurs pas seulement un processus mental ; elle est presque mécaniquement redoublée d'un clic qui l'opérationnalise instantanément et déclenche la mise en œuvre de l'action correspondante. Toujours dans l'exemple de Netflix, lorsque je clique sur un des films qui m'est recommandé, j'opérationnalise ma décision de le regarder, ce qui déclenche mon visionnage. À proprement parler, le système ne prend pas de décision (si ce n'est au sens 1) de détermination computationnelle), mais prend en charge toutes les conditions matérielles de possibilité de la décision, de sorte que l'opérationnalisation des processus mentaux de l'utilisateur peut être instantanée. Il préserve ainsi la capacité décisionnaire de l'utilisateur, dont le rôle est cependant réduit à cette portion congrue. Les connaissances infra-conceptuelles produites par la recommandation sont toujours déjà actualisables par l'interface, c'est-à-dire actionnables, et actualisées par le visionnage.

En revanche, dans la configuration en récit, l'articulation entre connaissance (encore en puissance), et action, est plus lointaine. Un récit transmis peut tout à fait rester lettre morte, n'être suivi d'aucune décision. Il peut tout aussi bien s'entendre comme un récit véridictionnel qui vise le vrai, que comme un récit pragmatiste visant l'action. Dans l'hypothèse c) où c'est la mise en œuvre des actions décidées sur la base des recommandations issues des conclusions de l'analyse de données qui valide et rend effectives ces conclusions, il y a une forte probabilité que les connaissances issues d'une configuration en récit ne soient jamais actualisées et demeurent seulement des connaissables possibles, comme les données elles-mêmes avant analyse. Quoi que l'analyste ait produit, le destinataire peut ne jamais prendre de décision sur la base de ces informations, ni mettre à exécution sa décision éventuelle.

De plus, les recommandations pourraient être évaluées non pas sur le fait que le décisionnaire a ou non mis en œuvre une action (qui confirmerait factuellement l'actionnabilité de ces recommandations) mais sur les conséquences pour l'entreprise de cette action. Ainsi, dans l'exemple cité plus haut, où l'analyste suggère de proposer un geste commercial aux clients qui ont subi une interruption de service, cette pratique pourrait avoir un coût trop élevé, nuisant à la rentabilité de l'entreprise dans l'équation économique entre la dépense correspondante à la somme de ces gestes commerciaux, et les bénéfices de l'amélioration de la satisfaction des clients. Dans cette hypothèse, les connaissances ne seraient pas évaluées par leur actionnabilité, mais par les conséquences économiques de leur activation. On notera d'ailleurs que cette suggestion ne peut pas être activée en tant que telle : il faut avant cela déterminer précisément la nature et le montant de ce geste commercial, et les caractéristiques de l'interruption, dont sa durée, si elle est totale ou partielle, etc. Le décisionnaire pourra également intégrer d'autres critères, comme la valeur du client estimée par un score interne : l'entreprise sera davantage encline à indemniser un client fidèle, qui représente un chiffre d'affaires supérieur à la moyenne, qui est susceptible de changer de fournisseur, etc. De cette façon, l'actionnabilité de la recommandation n'est plus de la seule responsabilité de l'analyste, mais elle devient également celle du décisionnaire, qui doit formaliser les modalités de son activation.

Idéalement, cette formalisation intègre également les conséquences économiques de ces modalités, à travers une projection économique qui permet de valider que cette activation est viable pour l'entreprise. L'ensemble des éléments produits constitue un plan d'action, qui explicite la décision de manière à la rendre effectivement actionnable. Ce plan d'action contient des connaissances pratiques, qui renvoient non pas à une représentation du monde, mais à des normes d'action.

La validité des connaissances produites est donc une responsabilité distribuée entre l'analyste et le décisionnaire destinataire du récit[57], dans un contexte où « rendre compte » des connaissances a une signification épistémique aussi bien qu'économique. Le plan d'action complète la recommandation de l'analyste pour former des connaissances actionnables. On peut parler d'une épistémologie sociale, c'est-à-dire d'une configuration dans laquelle la production de connaissances ne résulte pas d'un individu singulier, mais d'un ensemble d'agents qui contribuent, de la même façon ou de manière différente, à la production de connaissances. Ces agents collaborent successivement et/ou simultanément, pour arriver au résultat final. Le récit pris de manière autonome n'est qu'une partie de l'ensemble qui constitue une connaissable actionnable, à laquelle participe donc également le décisionnaire.

De ce fait, l'entreprise n'est pas un destinataire passif des connaissances produites par son prestataire. Par l'intermédiaire du décisionnaire (ou des personnes qui travaillent avec ou pour lui), elle est engagée dans le processus de production de connaissances. C'est du moins la configuration requise pour parvenir à des connaissances effectivement actionnables. En pratique, cette épistémologie sociale n'est pas forcément évidente, ni pour le commanditaire, ni pour le prestataire. Traditionnellement, le commanditaire d'une étude ne se voit pas comme son exécutant : son rôle consiste à exprimer ses attentes auprès du prestataire qui se charge de les combler moyennant finance. De même, le prestataire ne se considère pas comme responsable des conséquences économiques de l'activation de ses recommandations. L'épistémologie sociale des sciences des données productrices de connaissables actionnables reconfigure donc la nature des connaissances produites (par rapport à un institut d'études par exemple, qui produirait des conclusions sans recommandation) mais aussi les modalités organisationnelles de cette production. En ce sens, la production de connaissances actionnables est transformatrice des modes d'organisation et de fonctionnement : pour aboutir, le processus de production doit avoir des conséquences opérationnelles sur l'entreprise pour laquelle il a lieu. Ainsi, non seulement les connaissances produites doivent avoir des conséquences pratiques pour être validées pragmatiquement, mais leur processus de constitution lui-même doit avoir des conséquences similaires.

[57] Soulignons toutefois que le commanditaire de l'étude n'est pas forcément décisionnaire, qu'il n'est pas toujours évident de déterminer qui est décisionnaire par rapport à une problématique, ni même si un tel décisionnaire existe au sein de l'entreprise. Le quotidien d'une entreprise est aussi fait de ces « trous » dans les processus de décision qui freinent la mise en œuvre d'actions concrètes.

Néanmoins, ces modes de production et de validation restent à la fois nouveaux pour les entreprises, et difficiles à mettre en place. Une validation des connaissances qui prendrait la forme d'un jugement binaire, positif ou négatif, à partir des conséquences économiques des actions proposées par ces conséquences, ne pourrait que rarement être réalisée, et se conclurait le plus souvent par une invalidation du travail réalisé. De plus, les textes pragmatistes ne permettent pas, à notre connaissance, de trancher entre les différentes hypothèses de validation de connaissances que nous avons formulées, de la possibilité d'action à l'évaluation des connaissances économiques, en passant par la traduction de la décision en plan d'action, et par sa mise en œuvre effective : toutes ces hypothèses décrivent des connaissances qui ont des connaissances pratiques. Nous proposons donc de rejeter le recours à un critère de jugement binaire à partir des conséquences économiques de la mise en œuvre de la décision suggérée, pour adopter un système de validation graduelle des connaissances. Dans ce système, la connaissance n'est pas « utile » ou « inutile » de façon binaire et exclusive, mais présente un degré de validité qui peut être estimé par l'analyste et le commanditaire. Une telle estimation examine par exemple les critères suivants :

- la conclusion est présentée comme une proposition d'action ;
- la proposition d'action est explicite, précise, bien définie ;
- la proposition d'action mobilise des connaissances externes qui ne sont pas issues du traitement de données ;
- l'action proposée est en phase avec la stratégie de l'entreprise ;
- l'action proposée peut être traduite en un plan d'action ;
- le plan d'action est mis à exécution ;
- l'exécution du plan est concluante d'un point de vue économique.

Ainsi, dans notre exemple initial à partir de la conclusion

> « Les interruptions de service sont une source importante d'insatisfaction pour les clients. »

plusieurs propositions d'action peuvent être formulées, à commencer par celle de faire en sorte qu'il n'y ait pas d'interruption de service. Or la connaissance du secteur d'activité du client (en l'occurrence, un fournisseur d'accès à Internet) indique qu'il n'est pas possible de garantir une absence complète d'interruption de service (autrement, une coupure Internet), et qu'une réduction de ces interruptions aura un coût très élevé, nécessitant une révision majeure de l'infrastructure du réseau, mettant en jeu des parties prenantes externes qui devront elles aussi être convaincues de l'intérêt de cette révision. De ce fait, proposer un geste commercial est une solution qui apparaît d'emblée comme moins coûteuse. En revanche, elle suppose de développer une casuistique précise décrivant quel geste commercial doit être proposé en fonction de l'interruption rencontrée par le client, de sa portée, de sa durée, etc. Elle suppose aussi de pouvoir vérifier qu'il y a effectivement interruption, au-delà de la réclamation formulée par le client, ou d'opter le versement d'une remise sur la seule foi du client, sans

une vérification qui peut aussi avoir un coût de mise en place. Elle suppose également un plan d'action détaillant comment les chargés de relation client qui traitent les réclamations peut effectuer ce geste, et comment les former à cette nouvelle politique. Enfin, l'évaluation de cette politique de geste commercial doit elle-même être détaillée par la mise en point de critères d'évaluation et d'un plan de mise en œuvre de la mesure de ces critères. Bref, une proposition d'action entraîne avec elle une infinité de « détails » de mise en œuvre nécessaires pour que la proposition d'action devienne effective et qu'elle puisse être validée économiquement.

Face à cette complexité, le système de validité graduelle vise à déterminer non pas si la proposition est valide, mais à quel point elle l'est, ou autrement dit, jusqu'où le commanditaire aura pu aller dans sa mise en œuvre et sa validation. Bien qu'il ne repose pas sur une posture réaliste, il procède de normes similaires aux épistémologies confirmationnistes comme celle de Carnap, où la connaissance est validée par degré. Les connaissances produites procèdent d'une théorie qui a une structure confirmationniste à la Carnap, mais des normes pragmatistes, dans la mesure où le degré de validité des connaissances est principalement un degré d'actionnabilité, ou autrement dit, de « pragmaticité », des connaissances produites.

Nous avons ainsi vu que dans la configuration en système, les connaissances produites par le moteur de recommandation étaient toujours déjà des connaissances pratiques, qui ne visaient pas à décrire le réel, mais qui étaient des connaissances en tant qu'elles permettent une décision d'action (bien qu'il soit toujours possible de les réinscrire comme des connaissances infra-conceptuelles qui peuvent être ramenée à des concepts théoriques par un agent qui n'est pas l'utilisateur prévu du système). Du côté de la configuration en récit, cette relation entre connaissance et décision était moins évidente : en examinant les différentes conséquences pratiques que peut avoir une recommandation, nous avons proposé un système de validation graduelle de connaissances, fondé sur le degré d'actionnabilité, c'est-à-dire de pragmaticité, des connaissances produites. Néanmoins, il nous reste à examiner un point qui n'est pas couvert par la façon dont nous avons examiné l'articulation entre connaissance et décision dans la configuration en récit et la configuration en système : les configurations intermédiaires qui exploitent les conclusions d'une configuration en récit ou les usages des systèmes pour produire une forme spécifique de système, le tableau de bord.

3. Décider, connaître et mesurer

Les différentes configurations de production de connaissances à partir de données massives suscitent des actions qui peuvent à leur tour être mesurées. Les sciences des données permettent non seulement de produire des connaissances qui peuvent avoir des conséquences pratiques, mais d'évaluer ces conséquences. En entreprise, l'objet emblématique de cette activité de mesure et de contrôle est le **tableau de bord**.

Un tableau de bord est la représentation d'un ou plusieurs indicateurs de performance *(key performance indicators*, ou *KPI)* définis en fonction des objectifs d'un plan d'action. Le tableau de bord agrège ces indicateurs en un lieu unique. Il n'est pas nécessairement lié à une forme précise ou à un support computationnel, mais il atteint un niveau de concrétisation supplémentaire, en tant qu'objet technique, en prenant la forme non plus d'un panneau d'affichage ou d'un document numérique circulant entre décideurs, mais d'un logiciel dédié à cet usage. Dans notre exemple de l'introduction de gestes commerciaux en cas d'interruption de service, l'objectif à mesurer à travers un indicateur serait la réduction de l'insatisfaction liée à ces problèmes. L'indicateur de performance associé serait l'évolution (dans le temps) de la satisfaction après la mise en œuvre d'une batterie de gestes commerciaux. Un indicateur plus fin matérialiserait la satisfaction uniquement sur les clients concernés par ces interruptions de service. La notion même de satisfaction n'étant pas en soi quelque chose de mesurable, elle doit donc être convertie en indicateur pour trouver sa place dans un tableau de bord. La méthode du *Net Promoter Score* (NPS) que nous avions présentée au chapitre VI n'est dans cette perspective rien d'autre qu'une méthode de conversion d'une notion floue (la satisfaction) en un indicateur quantitatif qui peut varier d'une population à l'autre et en fonction du temps.

Toute réductrice que soit cette conversion, elle présente des similarités conséquentes avec la pratique scientifique. La conversion d'un phénomène en mesure renvoie au travail de modélisation du réel que réalise le chercheur. Tout modèle est une conversion réductrice du réel, parce que sa fonction épistémique est précisément de ramener le divers complexe du réel à un objet manipulable et mesurable. Un modèle de cellule ne dit rien de la variabilité des cellules possibles et de l'agencement exact de ses parties, de ses accidents, de ses évolutions. Néanmoins il est, dans une perspective instrumentaliste, suffisant pour les expérimentations dans lesquelles il est mobilisé. La ligne de partage majeure entre le modèle scientifique et l'indicateur de performance n'est pas leur réductionnisme, mais se situe au niveau de leur mode de validation.

En effet, un modèle scientifique doit en principe être validé avant d'être utilisé. Nous avons vu notamment au chapitre V comment la simulation informatique peut notamment avoir une fonction de déduction computationnelle des conséquences du modèle de manière à évaluer sa correspondance au réel. Les dispositifs expérimentaux qui s'appuient sur un modèle permettent également de le tester : c'est d'ailleurs cette relation de preuve expérimentale que la simulation s'efforce de reproduire dans certains cas. L'expérimentation est à la fois ce qui est opposable au modèle posé de manière hypothétique, et ce qui mobilise ensuite le modèle validé pour accroître la connaissance du réel.

À l'inverse, la conversion de la satisfaction en indicateur proposée par la méthode NPS s'appuie majoritairement, à notre connaissance, sur un argument d'autorité. Les sociétés de conseil qui accompagnent des entreprises dans la mise en œuvre de cette méthode mettent en avant le nombre et la diversité de leurs expériences passées ainsi que le caractère de norme qu'acquiert progressivement la méthode : les arguments reposent sur l'abondance d'autres entreprises ayant mis

en place cette méthode et le statut de standard qui lui est conféré par cette abondance. Elles mettent notamment en avant des exemples d'entreprises dont la réussite économique est reconnue, suggérant une corrélation entre cette réussite et la mise en place de la méthode NPS qui est implicitement envisagée comme une relation de cause à effet. Ainsi, s'il y a justification de la modélisation dont résulte l'indicateur, ce n'est encore une fois pas par correspondance au réel, mais *via* les conséquences pratiques de la mise en œuvre de l'indicateur. De plus, comme l'activité de modélisation scientifique, le travail de conception des indicateurs n'est en soi pas infaillible ; un indicateur peut n'avoir que l'apparence d'une correspondance au réel, et ne constituer qu'une forme sans contenu. Il y a des bons et des mauvais indicateurs, comme il y a de bons et de mauvais modèles, par rapport à un contexte ou une finalité donnée.

Malgré cette différence, dans la justification, par rapport aux connaissances scientifiques, un indicateur de performance renvoie toutefois à une notion plus classique de la connaissance que celle qui est à l'œuvre dans la recommandation. Un indicateur de performance prend la forme d'une mesure qui permet de savoir quelque chose à propos de quelque chose d'autre. Il apparaît donc dans un régime de correspondance au réel par lequel un indicateur de satisfaction est fondé sur des données qui ont un statut de mesure de la satisfaction des clients, et non d'une satisfaction comme notion abstraite décrivant seulement les données. À travers cet indicateur, la donnée est chargée rétroactivement d'une responsabilité épistémique de description du réel.

À ce titre, les indicateurs de performance fondés sur l'exploitation des conséquences pratiques des recommandations issues du traitement de données massives renvoient à des pratiques plus anciennes de contrôle de l'activité en entreprise. La notion d'indicateur de performance remonte en effet aux pratiques de *benchmarking* issues d'innovations gestionnaires et organisationnelles mises au point à partir des années 1920 aux États-Unis et au Japon (Bruno et Didier 2013). À partir de théories du contrôle du travail et du management de la qualité, ces pratiques visent à mettre en compétition différents modes de travail au sein de l'entreprise ou d'une entreprise à l'autre. Cette mise en compétition repose sur la définition d'éléments de comparaison, les indicateurs, et vise à identifier les « meilleures pratiques » observables dans ou au dehors de l'entreprise, et rendues commensurables à travers les indicateurs qui les évaluent.

Le benchmarking rend possible la mise en compétition des services et des salariés de l'entreprise. Augmenté par la mobilisation des traces de l'activité de l'entreprise et de ces clients, ce benchmarking peut être automatisé en agrégeant et en visualisant les données qui alimentent ces indicateurs de performance. Dans la représentation de données, les visualisations de données jouent un rôle de description et de preuve similaire à celui qu'elles occupent dans les configurations en récit. Les graphiques d'un récit et d'un tableau de bord peuvent d'ailleurs être identiques du point de vue de leur forme comme de leur contenu. Un graphique montrera l'évolution de la satisfaction, l'autre le nombre de réclamations traitées, un troisième les principaux motifs qui suscitent une réclamation,

etc. Toujours dans l'exemple de la gestion des interruptions de service, la satisfaction des clients à l'issue du traitement de leur réclamation peut être mesurée en gardant la trace du chargé de clientèle qui a traité la réclamation. En agrégeant ces mesures, le tableau de bord permet non seulement de comprendre les facteurs d'amélioration de la satisfaction client, mais aussi de repérer nominalement les chargés de clientèle dont les actions suscitent un niveau de satisfaction inférieur à la moyenne.[58] De manière générale, c'est la performance globale de l'entreprise qui doit être améliorée en améliorant en moyenne celle des individus, et non en ciblant spécifiquement les individus jugés moins performants.

Le benchmarking comme technique fondée sur une correspondance au réel des indicateurs reste en effet au service de finalités pratiques. La responsabilité épistémique de description du réel que possèdent les indicateurs est à mettre en regard de leur fonction d'instruments d'évaluation des conséquences économiques de plans d'actions mis en œuvre (ou non) sur la base de recommandations issues du traitement des traces d'activité de l'entreprise. Tandis que les recommandations aident les décideurs à définir des plans d'action, les indicateurs évaluent rétrospectivement ces recommandations à travers la mesure des conséquences des actions mises en œuvre à partir de ces recommandations. De plus, ce processus n'est pas uniquement linéaire, de la recommandation au tableau de bord : ce dernier peut lui aussi suggérer des ajustements du plan d'action, voire une nouvelle stratégie, ce qui en fait également une source de recommandations.

Des tableaux de bord peuvent exister à la fois à la suite de projets correspondant une configuration en récit, et dans le cadre de la conception de systèmes de recommandation. Dans le premier cas, les conclusions du récit, en mettant en évidence les principaux enseignements que l'on peut tirer des données, contribuent directement à la définition des indicateurs de performance. Le récit identifie les points clés tandis que le tableau de bord va permettre de suivre leur évolution dans le temps. On peut également voir le tableau de bord comme un schéma narratif spatialisé et dynamique, avec lequel on interagit pour identifier des relations entre variables, et dont l'utilisateur construit le sens et le « dénouement » *via* ces interactions. Ce schéma fondé sur une synthèse spatiale s'apparente par conséquent à un système tel que nous l'avons décrit.

Du côté des configurations en système à proprement parler, la plupart des logiciels fondés sur un traitement de données significatif intègre des fonctionnalités de suivi de l'activité des utilisateurs, qui sert à la fois à enrichir rétroactivement la recommandation, mais aussi, et dans ce but, à suivre et mieux comprendre les comportements des utilisateurs. L'analyse de ce comportement ne se fait pas par la consultation des journaux d'utilisation *(logs)* bruts, mais à travers des vues agrégées dans des tableaux de bord. Dans les échanges que nous avons eus avec des spécialistes de la recommandation

[58] Cette forme de benchmarking permet un contrôle individualisé de l'activité professionnelle. A notre connaissance, cette forme de surveillance individuelle n'est cependant pas pratiquée, du moins chez les clients de Proxem, à l'exception d'un projet qui a permis de retracer l'auteur de pratiques illégales.

algorithmique, la mesure de la qualité des recommandations à travers des indicateurs de performances fondés sur la mesure du comportement des utilisateurs face à ces recommandations fait partie du quotidien. Ces comportements utilisateurs sont d'abord analysés de manière à en tirer des indicateurs de performance comme dans une configuration en récit, puis intégrés à un tableau de bord qui permet d'en visualiser l'évolution. Un tableau de bord est donc un objet compatible avec de multiples situations, qu'il s'agisse de traitements de données massives dans le cadre d'une configuration en récit ou en système, ou d'autres manières d'alimenter des indicateurs de performance au sein de l'entreprise.

Il convient de souligner que les décisions prises en entreprise ne procèdent pas systématiquement de recommandations issues de l'analyse des traces de leur activité, et toutes ne sont pas évaluées à travers des indicateurs de performance. Les entreprises qui s'appuient sur les mêmes données pour améliorer leurs prises de décision et le contrôle de leur activité restent rares, car ces pratiques n'ont rien de nécessaire du point de vue du fonctionnement de l'entreprise : l'antériorité des pratiques de benchmarking sur les sciences des données atteste du caractère non-nécessaire de ces dernières, sans lesquelles les premières peuvent tout à fait exister. Néanmoins, certaines entreprises adoptent ces nouvelles pratiques, du fait qu'elles renvoient à des enjeux concrets que sont le pilotage et le contrôle des entreprises, et qu'elles peuvent apporter une amélioration dans la concrétisation de ces enjeux. De ce point de vue, les recommandations issues des sciences des données ont un double rôle de *légitimation* des plans d'action mis en œuvre par les décideurs au regard de normes empiristes qui rejettent des décisions d'action intuitives fondées sur l'intuition individuelle, et un rôle de production partielle (sur le mode de la suggestion) des décisions qui donnent lieu à ces plans d'action. On retrouve cette structure dans la configuration en système où la recommandation auto-justifie la décision qu'elle propose, et la rend effectuable par l'utilisateur. Pour leur part, les tableaux de bord ont également une double fonction 1) de contrôle par la mesure de l'activité de l'entreprise, et par leur intermédiaire, des décisions qui gouvernent cette activité, et 2) de production itérative de nouvelles décisions permettant d'ajuster ou de réorienter les plans d'actions qui régissent, au moins en principe, l'activité des entreprises.

En synthèse, dans le cadre pragmatiste, les sciences des données engendrent et mesurent des conséquences pratiques, au moyen de connaissances pragmatistes. Ces connaissances sont évaluées à l'aune de leurs conséquences pratiques, produites par un ensemble de décisions individuelles ou collectives et de déterminations computationnelles, et constituent elles-mêmes des propositions d'action, c'est-à-dire des décisions en puissance actualisées par la mise en œuvre de plans d'action correspondants.

Conclusion

Au terme de cette investigation dans les discours et les pratiques d'analyse de données massives, nous avons mis au jour un paradigme méthodologique dans lequel pourront s'inscrire des projets futurs d'exploitation des traces numériques. Ce paradigme régit comment une valeur épistémique est attribuée à la donnée, mais aussi comment construire et manipuler des corpus, comment interpréter les résultats du calcul, comment restituer l'investigation de manière à la rendre intelligible pour un lecteur ou utilisateur, et comment déterminer la validité des connaissances produites. Il met en évidence le caractère déterminant de la notion de décision qui régit l'ensemble du processus ainsi que l'articulation entre connaissance et action. Deux composantes se révèlent également essentielles à l'intégration de ce paradigme : le rôle de l'**artisanat** dans la constitution en donnée, leur préparation, leur nettoyage et leur analyse, et celui de la **rhétorique** de l'image et du récit, crucial pour faire émerger et rendre intelligible le sens des données. Pour cela, plusieurs concepts clés sont proposés :

- La distinction entre donnée et **trace** fait de celle-ci un signe toujours intermédiaire et indirect du réel visé, dont le sens est à construire en situation par extraction d'un corpus. La **valeur épistémique** de la trace numérique est le fruit d'une construction au sein d'un **paradigme indiciaire**.
- La distinction entre **discipline à terrain** et **discipline à corpus** affine une seconde fois l'ancrage de ces pratiques dans les sciences de la culture. Elle permet de préciser un accès au réel qui se fait en substituant à une représentativité statistique de la donnée une **logique de constitution** du corpus dont le récit rend raison. Cette logique de constitution, qui permet de décider de ce qui doit être inclus ou exclu, procède d'un **projet épistémique** qui prend forme progressivement au fil de l'**investigation,** et ne préexiste pas aux données comme le feraient des hypothèses de recherche dans le cadre de l'analyse de données classique.
- Cette constitution épistémique du corpus se double d'une **constitution technique**, au sein de laquelle la trace numérique est toujours déjà **manipulable** du fait de son support informatique, et doit être nettoyée et formatée dans un travail guidé par ce projet épistémique, et qui appartient déjà à la phase d'analyse de données. L'informatique joue ainsi le rôle d'un

transcendantal empirique qui détermine *a posteriori*, de manière effective, les conditions de possibilité des connaissances à produire.

- L'analyse de données s'inscrit dans un **paradigme exploratoire**, qui, contrairement à son pendant confirmatoire, s'efforce de donner à voir et rendre raison des données elles-mêmes qui constituent pour l'analyste le réel à connaître. Elle mobilise la boîte à outil des **sciences des données** envisagées comme agrégat de techniques provenant de lignées différentes, capable de circuler entre projets d'analyse car détachées de leur ancrage théorique à travers la suppression de l'antériorité de la modélisation sur le calcul.
- En l'absence de cette étape de modélisation, la manipulation computationnelle effectuée par les sciences des données est nécessaire mais pas suffisante pour rendre raison du corpus et faire émerger des significations. C'est l'intermédiaire de l'interprétation et de la **visualisation de données** qui permet de rendre intelligibles les résultats du calcul tout en endossant un rôle de preuve de ces résultats.
- Ces résultats peuvent être restitués selon deux formes. La première est un **récit** qui réalise une synthèse de l'hétérogène et dont la narration permet de rendre raison de l'investigation. Cette narration est constituée de composantes visuelles agrémentées de contextualisations textuelles qui matérialisent les connaissances et en stabilisent l'interprétation.
- La seconde forme est le **système**, à travers le principe du logiciel de **recommandation** qui découpe l'analyse de données en un travail d'investigation et un travail d'industrialisation qui rend cette investigation répétable. La matérialisation de suggestions à travers une interface propose à la figure de l'utilisateur des interprétations entérinées par ses interactions avec le logiciel.
- Quelle que soit la forme de restitution, l'analyse de traces numériques promeut un empirisme rationnel qui fonde son accès au réel dans la trace et mobilise le calcul, la visualisation et le récit comme justification de ses raisonnements. À chaque étape, l'exercice d'**auto-réfutation** met à l'épreuve la rationalité de l'investigation, et la fonde en raison lorsque les tentatives d'auto-réfutation sont sans succès.
- À chaque instant, cette investigation procède de **décisions** humaines, de choix de conceptualisation et de conception face auxquels le calcul n'a jamais un caractère définitif, mais sert seulement de support et d'aide à la production de connaissances à partir des traces.
- Enfin, cet empirisme rationnel est mobilisé dans une rhétorique de la scientificité qui facilite la confiance dans les connaissances produites, mais n'en constitue pas le régime de validité : c'est en fin de compte l'**actionnabilité** des connaissances, c'est-à-dire leur capacité à suggérer des décisions d'action pertinentes pour le destinataire de l'investigation, qui détermine leur validité à partir des conséquences des actions qu'elles suggèrent.

Le paradigme méthodologique ainsi défini rend possible des sciences computationnelles de la culture qui mobilisent à la fois la culture épistémique des sciences de la culture, le savoir-faire technique des sciences des données, et une approche méthodologique spécifique articulée autour de l'exploration et de la visualisation. Ces sciences computationnelles de la culture ont ici encore un statut largement hypothétique, avec des opérationnalisations partielles et très localisées. En particulier, l'appareil conceptuel des sciences de la culture y est encore largement sous-exploité, du fait de l'ancrage institutionnel des pratiques analysées. Ces pratiques se jouent loin des sciences de la culture, en entreprise, dans la recherche en informatique ou encore chez des analystes amateurs ou autodidactes.

De ce point de vue, une inquiétude et un espoir nous semblent devoir être formulées quant au devenir de ce paradigme méthodologique.

Premièrement, l'exploitation des traces numériques par la recherche et l'industrie informatique se caractérise par une sous-mobilisation de l'appareil conceptuel des sciences de la culture, généralement ignoré des praticiens du calcul, et remplacé par des concepts rudimentaires issus du sens commun, de plus en plus problématiques d'un point de vue épistémologique et éthique. La naturalisation des objets des sciences de la culture, telles que la criminalité (Xiaolin Wu et Zhang 2016) ou l'orientation sexuelle (Wang et Kosinski 2017) revient à un retour à la phrénologie et à une négation sommaire des résultats des sciences de la culture dans leur ensemble. Ces travaux ont un caractère performatif par lequel ils nourrissent les dystopies technologiques dont ils s'inspirent, et ignorent la responsabilité éthique et sociale qui doit imprégner la recherche sur l'homme et la société. Il y a ainsi une certaine urgence, non seulement à invalider ces travaux, mais à revaloriser l'appareil conceptuel des sciences de la culture et sa capacité à interroger les conséquences de leur activité de recherche ; bref, il est de plus en plus nécessaire de montrer qu'une autre voie est possible.

Deuxièmement, les sciences de la culture et l'industrie informatique n'ont pas de nécessité à entrer en guerre. Les capacités offertes par le calcul sont à voir comme une chance pour notre connaissance du fait humain, et doivent pouvoir fonctionner sous un régime de complémentarité plus que d'opposition. La voie de l'interdisciplinarité entre des cultures épistémiques foncièrement différentes n'a certes rien de simple ou de rapide, et nécessite d'être prêt à fabriquer un tout qui sera plus que la somme de ses parties en faisant le deuil de la pureté méthodologique et théorique de chacune d'elles. À l'échelle du collectif comme de l'individu, elle nécessite un compromis entre des conceptions du monde hétérogènes et trop souvent vues comme incompatibles. Elle exige d'être un peu philosophe, d'accepter de désapprendre ce que l'on sait, et comment on le sait. Néanmoins, comme nous l'avons illustré avec notre étude de cas, une telle alliance est possible, et chaque camp a à y gagner, non seulement en termes de volume et de validité des connaissances qu'il peut en tirer, mais en termes d'enrichissement de lui-même.

Cette thèse plaide pour le métissage, privilégiant l'espoir à l'inquiétude. Iréniste, elle s'est efforcée de ne pas s'inscrire dans un camp, de ne pas même en établir, mais de dessiner des composantes complémentaires, et de rendre compte de ce que chacune avait à apporter. Formellement, sa contribution est une démarche générale pour articuler ces composantes. Dans sa continuité, l'exploration et la compréhension des conditions de possibilité de ce métissage, à travers la démarche proposée comme par d'autres moyens, est le programme de recherche que nous nous donnons à présent.

Glossaire conceptuel

cartographie : suivant notre définition, forme de visualisation adaptée à l'exploration de grandes masses de données, visant non pas à montrer visuellement et quantitativement une idée spécifique, mais à représenter un espace de sens défini par des données. Le dispositif cartographique permet de dégager des formes dans cet espace. Globale, synoptique et interactive, la représentation ou carte devient le territoire à explorer. Le graphe, la carte géographique, la carte de chaleur, sont des exemples de cartographies.

critique : on distingue traditionnellement trois formes de critiques :

- la critique ou analyse littéraire, qui vise à mettre en évidence le sens et la spécificité d'un texte ;
- la critique kantienne, qui vise à déterminer les frontières, et ainsi délimiter le domaine d'un concept ;
- la critique sociale, qui vise à dénoncer un fait ou un propos de manière à en montrer l'illégitimité.

compréhensif : en épistémologie des sciences de la culture, le régime compréhensif s'oppose au régime explicatif en ceci qu'il recherche des connaissances fondées non pas sur la mise en évidence de causes relatives à des phénomènes envisagés d'un point de vue externaliste, mais sur l'identification des raisons de phénomènes analysés d'un point de vue internaliste et empathique, de manière à *comprendre* les motivations de ces phénomènes. Chez Weber, il s'agit de reconstituer le sens que les acteurs donnent à l'action, et chez Dilthey, de retrouver le vécu correspondant à un terme ou une situation.

configuration en récit : suivant notre définition, forme de restitution (d'une analyse de données) appartenant au genre narratif, qui mobilise la cohérence d'un récit pour rendre raison de l'analyse de données et de ses résultats, et s'appuie sur une enquête unique.

configuration en système : suivant notre définition, forme de restitution (d'une analyse de données) mobilisant l'industrie logicielle, qui consiste à rendre répétable le processus d'analyse de données de

manière à créer les conditions de possibilité d'une décision prenant la forme d'un choix entre plusieurs suggestions.

digital humanities : d'après notre analyse, champ de recherche anglo-saxon qui problématise, conceptualise et expérimente sur les transformation du texte à l'heure du numérique. Voir aussi **humanités numériques**.

discipline à corpus : suivant notre définition, branche des sciences de la culture dont l'accès au réel passe par la médiation du corpus comme ensemble de documents textuels, ou non, qui constitue des traces de l'événement. Y figurent notamment l'histoire, l'analyse littéraire, la linguistique. Ces disciplines sont généralement idiographiques. Voir aussi **disciplines à terrain**.

discipline à terrain : suivant notre définition, branche des sciences de la culture dont l'accès au réel est le terrain, soit par l'observation directe, soit par la médiation de l'enquête et de la statistique inférentielle. Ces disciplines produisent des données qui peuvent être analysées suivant un régime idiographique mais aussi nomologique, d'une manière similaire aux sciences de la nature. Y figurent notamment la sociologie, l'anthropologie, la géographie. Voir aussi **disciplines à corpus**.

décision : suivant notre définition, une décision est une action de l'esprit qui consiste, au terme d'un processus de délibération, à opter pour un choix entre différentes possibilités. Il ne s'agit pas d'un comportement observé de l'extérieur, mais d'un processus interne qui n'a pas forcément de conséquences observables. La décision désigne à la fois le processus de délibération et sélection, et le résultat de ce processus.

exploration de données : chez John Tukey notamment, ensemble d'actions et d'heuristiques s'appuyant sur le calcul et la visualisation et visant prendre connaissance d'un ensemble de données, à le décrire, à former et à tester des hypothèses relativement à cet ensemble ou une sous-partie de celui-ci.

interprétatif : en épistémologie des sciences de la culture, régime de connaissance où la signification des observables ne procède pas d'un processus mécanique tel que le calcul mais d'un travail d'analyse, l'interprétation, qui propose une signification à partir de signes à interpréter, de connaissance externes, et d'une tradition d'interprétation ancrée dans une culture épistémique.

humanités numériques : d'après notre analyse, champ de recherche francophone qui interroge, et, secondairement, expérimente, sur l'articulation entre les technologies numériques et les sciences humaines et sociales. Voir aussi **digital humanities**.

organon : en philosophie des sciences, ensemble d'outil techniques et méthodologiques permettant d'opérationnaliser l'activité épistémique.

paradigme méthodologique : dans cette thèse, ensemble de propositions méthodologiques régies par un cadre conceptuel cohérent, visant à rendre possible la production de connaissances à partir de traces numériques, et à assurer la validité de ces connaissances.

paradigme indiciaire : en épistémologie des sciences de la culture, cadre conceptuel au sein duquel les connaissances sont produites par le biais d'une enquête qui s'appuie sur l'interprétation de signes (ou indices) jouant le rôle d'une représentation d'un phénomène absent.

projet épistémique : suivant notre définition, ensemble d'hypothèses relatives à l'articulation entre des données et des connaissances, qui prend forme progressivement à travers l'initiation d'une recherche, la constitution d'un corpus et l'exploration des données.

recommandation : suivant notre analyse, résultat généré par un système computationnel et matérialisé à travers une interface utilisateur. Une recommandation est une proposition, ou suggestion, de ce système qui restreint computationnellement les actions possibles en mettant en avant les plus pertinentes, tout en permettant un choix. La pertinence est définie et formalisée en amont par le concepteur du système. Par extension, la recommandation désigne la démarche requise pour mettre en place un système de recommandation.

sciences computationnelles de la culture : suivant la proposition de cette thèse, ensemble de normes et de pratiques hypothétique visant à faire émerger, à travers l'analyse computationnelle de traces numériques, de nouvelles connaissances relatives au fait humain, qu'il soit social, politique ou culturel.

sciences de la culture : en philosophie des sciences, versant de la connaissance scientifique caractérisé par son approche (idiographique) et son objet (le fait humain). Voir aussi **sciences de la nature**.

sciences de la nature : en philosophie des sciences, versant de la connaissance scientifique caractérisé par son approche (nomologique) et son objet (la nature). Voir aussi **sciences de la culture**.

suggestion : suivant notre analyse, une suggestion est dans le cadre d'un processus de décision une présélection préalable d'options dans un espace de possible. Une recommandation est une forme de suggestion à laquelle on attribue une certaine pertinence.

trace : résidu matériel d'un événement toujours déjà passé, qui reçoit une signification en étant détaché de son contexte ou environnement. Ce détachement ou singularisation prend sens par rapport à une interprétation qui dépend elle-même d'un lecteur et de son régime interprétatif. Par le lien qu'elle entretient avec l'événement, la trace peut accéder à un statut de représentation de celui-ci.

transcendantal empirique : formule oxymorique présente chez quelques postkantiens comme Cassirer, ici mobilisée pour désigner le statut joué par l'informatique et les sciences des données dans

la production de connaissances à partir de traces numériques, en déterminant empiriquement les conditions de possibilité, le domaine et les limites de ces connaissances.

Bibliographie

ADAM Matthias, 2004, « Why worry about theory-dependence? Circularity, minimal empiricality and reliability », *International Studies in the Philosophy of Science*, 2004, vol. 18, nº 2-3, p. 117-132.

AGRAWAL Rakesh et SRIKANT Ramakrishnan, 1994, « Fast algorithms for mining association rules », *Proceedings of the 20th International Conference on Very Large Data Bases*, 1994.

AMOSSE Thomas, 2013, « La nomenclature socio-professionnelle : une histoire revisitée », *Annales. Histoire, Sciences Sociales*, 2013, nº 4, p. 1039-1075.

ANDERSON Chris, 2008, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory , 2008, consulté le 11 mai 2014.

ANDLER Daniel, 2011, « Le naturalisme est-il l'horizon scientifique des sciences sociales ? » dans *Les sciences humaines sont-elles des sciences ?*, Paris, Vuibert, p. 15-34.

ANDRADE N, 2010, « Technology and Metaphors: from Cyberspace to Ambient Intelligence », *Observatorio (OBS)*, 2010, vol. 4, p. 121-146.

ANSCOMBE F J, 1973, « Graphs in statistical analysis », *The American Statistician*, 1973, vol. 27, nº 1, p. 441-475.

ARMATTE Michel, 2008, « Histoire et Préhistoire de l'Analyse des données par J.P. Benzecri : un cas de généalogie rétrospective », *Electronic Journ@l for History of Probability and Statistics*, 2008, vol. 4, December, p. 1-24.

ARMSTRONG David Mallet, 1993, « A World of States of Affairs », *Philosophical Perspectives*, 1993, vol. 7, Language and Logic, p. 429-440.

ASUR Sitaram et HUBERMAN Bernardo A., 2010, « Predicting the Future with Social Media », Toronto, IEEE.

BACHIMONT Bruno, 2010a, « La présence de l'archive : réinventer et justifier », *Intellectica*, 2010, vol. 54, p. 281-309.

BACHIMONT Bruno, 2010b, « Le numérique comme support de la connaissance : entre matérialisation et interprétation » dans Ghislaine Gueudet et Luc Trouche (eds.), *Ressources vives. Le travail documentaire des professeurs en mathématiques*, Rennes, Presses universitaires de Rennes, p. 1-12.

BACHIMONT Bruno, 2010c, *Le sens de la technique : Le numérique et le calcul*, Paris, Encre Marine.

BACHIMONT Bruno, 1996, *Herméneutique matérielle et artéfacture, Des machines qui pensent aux machines qui donnent à penser*,Ecole Polytechnique, Thèse de doctorat en épistémologie, 354 p.

BARABÁSI Albert-László, 2013, « Network science », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 18 février 2013, vol. 371, nº 1987, p. 20120375-20120375.

BARABÁSI Albert-László, 1999, « Emergence of Scaling in Random Networks », *Science*, 15 octobre 1999, vol. 286, nº 5439, p. 509-512.

BARBEROUSSE Anouk et IMBERT Cyrille, 2014, « Recurring models and sensitivity to computational constraints », *Monist*, 2014, vol. 97, nº 3, p. 259-279.

BARBEROUSSE Anouk et IMBERT Cyrille, 2013, « Le tournant computationnel et l'innovation théorique » dans Soazig Le Bihan (ed.), *Précis de philosophie de la physique*, Paris, Vuibert, p. 1-26.

BARBIN Évelyne. et LAMARCHE Jean-Pierre, 2004, *Histoires de probabilités et de statistiques*, Paris, Ellipses.

BARLOW John Perry, 1996, « A Declaration of the Independence of Cyberspace. », *Humanist*, 1996, vol. 56, nº 3, p. 18-19.

BARTHES Roland, *Mythologies*, Paris, Editions du Seuil.

BASTARD Irène, CARDON Dominique, FOUETILLOU Guilhem, PRIEUR Christophe et RAUX Stéphane, 2013, *Les sciences sociales et les données indiscrètes du web*, http://www.lemonde.fr/sciences/article/2013/12/16/les- sciences-sociales-et-les-donnees-indiscretes-du- web_4335257_1650684.html , 2013.

BASTIAN Mathieu, HEYMANN Sebastien et JACOMY Mathieu, 2009, « Gephi: An Open Source Software for Exploring and Manipulating Networks », *Third International AAAI Conference on Weblogs and Social Media*, 2009, p. 361-362.

BEAUD Jean-Pierre et PREVOST Jean-Guy, 2000, *L'ère du chiffre. Systèmes statistiques et traditions nationales*, Québec, Presses de l'Université du Québec.

BEER David, 2017, « The data analytics industry and the promises of real-time knowing: perpetuating and deploying a rationality of speed », *Journal of Cultural Economy*, 2 janvier 2017, vol. 10, nº 1, p.

21-33.

BENEL Aurélien, 2013, « Quelle interdisciplinarité pour les « humanités numériques » ? », *Les cahiers du numérique*, 2013, vol. 1, p. 1-23.

BENEL Aurélien, LEJEUNE Christophe et ZHOU Chao, 2010, « Eloge de l'hétérogénéité des structures d'analyse de textes », *Document numérique*, 2010, vol. 13, nº 2, p. 41-56.

BERNERS-LEE Tim, HALL Wendy, HENDLER James A, O'HARA Kieron, SHADBOLT Nigel et WEITZNER Daniel J, 2006, « A Framework for Web Science », *Foundations and Trends® in Web Science*, 2006, vol. 1, nº 1, p. 1-130.

BERRY David M., 2012, *Understanding Digital Humanities*, Londres, Palgrave Macmillan UK, 336 p.

BERTIN Jacques, 1970, « La graphique », *Communications*, 1970, vol. 15, p. 169-185.

BOLLEN Johan, MAO Huina et ZENG Xiaojun, 2011, « Twitter mood predicts the stock market », *Journal of Computational Science*, 14 mars 2011, vol. 2, nº 1, p. 1-8.

BOURGET David et CHALMERS David J, 2013, « What Do Philosophers Believe? », 2013, p. 1-37.

BOWKER Geoffrey C, 2014, « The Theory/Data Thing Commentary », 2014, vol. 8, nº 2043, p. 1795-1799.

BOWKER Geoffrey C. et STAR Susan Leigh, 1999, *Sorting things out. Classification and its consequences*, Cambridge, Massachusetts, The MIT Press, 389 p.

BOYD Danah et CRAWFORD Kate, 2011, « Six Provocations for Big Data », Oxford.

BOYER-KASSEM Thomas, 2014, « Layers of Models in Computer Simulations », *International Studies in the Philosophy of Science*, 2 octobre 2014, vol. 28, nº 4, p. 417-436.

BREIMAN Leo, COX DR et BREIMAN Leo, 2001, « Comment-Statistical Modeling: The Two Cultures », *Statistical Science*, 2001, vol. 16, nº 3, p. 199-231.

BROOKS Laura L. et OWEN Richard, 2008, *Answering the ultimate question : how Net Promoter can transform your business*, Chichester, John Wiley & Sons, 320 p.

BRUGGER N., 2012, « Historical Network Analysis of the Web », *Social Science Computer Review*, 6 septembre 2012, vol. 31, nº 3, p. 306-321.

BRUNO Isabelle et DIDIER Emmanuel, 2013, *Benchmarking : l'état sous pression statistique*, Paris, Zones.

BURDICK Anne, DRUCKER Johanna, LUNENFELD Peter, PRESNER Todd et SCHNAPP Jeffrey, 2012, *Digital Humanities*, Cambridge, Massachusetts, MIT Press.

CAILLIAU Frederik et POUDAT Céline, 2008, « Caractérisation lexicale des contributions clients agents

dans un corpus de conversations téléphoniques retranscrites », *JADT 2008 : 9e Journées internationales d'Analyse statistique des Données Textuelles*, 2008, p. 267-275.

CALUDE C.S. et LONGO G., 2015, « The Deluge of Spurious Correlations in Big Data », *CDMTCS Research Report Series*, 2015, August, p. 1-13.

CARDON Dominique, 2013, « Dans l'esprit du PageRank », *Réseaux*, 2013, vol. 177, nº 1, p. 63.

CARNINO Guillaume, 2010, « Les transformations de la technologie : du discours sur les techniques à la "techno-science" », *Romantisme*, 2010, vol. 150, nº 4, p. 75.

CASILLI Antonio, 2012, « Comment les usages numériques transforment-ils les sciences sociales? » dans Pierre Mounier (ed.), *Read/Write Book 2. Une introduction aux humanités numériques*, Marseille, OpenEdition Press, p. 1-9.

CAVENG Remy, 2012, « La production des enquêtes quantitatives », *Revue d'anthropologie des connaissances*, 2012, vol. 6, nº 1, p. 65.

CHALMERS Alan, 2003, « The Theory-Dependence of the Use of Instruments in Science », *Philosophy of Science*, 2003, vol. 70, nº 3, p. 493-509.

CHAMPIN Pierre-Antoine, MILLE Alain et PRIE Yannick, 2013, « Vers des traces numériques comme objets informatiques de premier niveau: une approche par les traces modélisées », *Intellectica*, 2013, vol. 59, nº 1, p. 1-33.

CHATEAURAYNAUD Francis, 2003, *Prospéro : une technologie littéraire pour les sciences humaines*, Paris, Editions CNRS.

CHAUMARTIN François-Régis, 2012, *Antelope, une plate-forme de TAL permettant d'extraire les sens du texte*,Paris Diderot (Paris 7), Thèse de doctorat en linguistique théorique, descriptive et automatique.

CIBOIS Philippe, 1990, « Eclairer le vocabulaire des questions ouvertes par les questions fermées: le tableau lexical des questions », *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 1990, vol. 26, nº 1, p. 12-21.

CIOFFI-REVILLA Claudio, 2014, *Introduction to Computational Social Science*, Londres, Springer-Verlag, 320 p.

CLEVELAND William S et MCGILL Robert, 1985, « Graphical perception and graphical methods for analyzing scientific data. », *Science (New York, N.Y.)*, 1985, vol. 229, nº 4716, p. 828-833.

COUPER Mick P., 2000, « Web surveys: A review of issues and approaches », *Public opinion quarterly*, 2000, vol. 64, nº 4, p. 464-494.

D'AUBIGNY Gérard, 2001, « Le traitement des questions ouvertes dans les enquêtes et sondages :

introduction », *Journal de la société statistique de Paris*, 2001, vol. 142, nº 4, p. 1-5.

DACOS Marin et MOUNIER Pierre, 2014, *Humanités numériques : état des lieux et positionnement de la recherche française dans le contexte international*, Rapport commandé par l'Institut français, opérateur du ministère des Affaires étrangères pour l'action culturelle extérieure de la France.

DASTON Lorraine J. et GALISON Peter Louis, 2007, *Objectivity*, Cambridge, Massachusetts, MIT Press, 501 p.

DEERWESTER Scott, DUMAIS Susan T., FURNAS George W., LANDAUER Thomas K. et HARSHMAN Richard, 1990, « Indexing by latent semantic analysis », *Journal of the American Society for Information Science*, 1990, vol. 41, nº 6, p. 391-407.

DENIS Jérôme et PONTILLE David, 2011, « Materiality, Maintenance and Fragility. The Care of Things », Corfou.

DESROSIERES Alain, 2010, *La politique des grands nombres. Histoire de la raison statistique*, Paris, La Découverte.

DIAKOPOULOS Nicholas, 2015, « Algorithmic Accountability », *Digital Journalism*, 2015, February 2015, p. 1-18.

DIOGO Maria Paula, 2017, « « To be or not to be ». La construction d'une identité civile pour les ingénieurs militaires portugais », *Quaderns d'Història de l'Enginyeria*, 2017, vol. 15, p. 243-256.

DODDS Peter Sheridan, CLARK Eric M., DESU Suma, FRANK Morgan R., REAGAN Andrew J., WILLIAMS Jake Ryland, MITCHELL Lewis, HARRIS Kameron Decker, KLOUMANN Isabel M., BAGROW James P., MEGERDOOMIAN Karine, MCMAHON Matthew T., TIVNAN Brian F. et DANFORTH Christopher M., 2015, « Human language reveals a universal positivity bias », *Proceedings of the National Academy of Sciences*, 2015, vol. 112, nº 8, p. 2389-2394.

DOMINGOS Pedro, 2015, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York, Basic Books, 354 p.

DOUEIHI Milad., 2008, *La grande conversion numérique*, Paris, Éditions du Seuil.

DOUVEN Igor, 2011, « Abduction » dans Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2017 Edition)*, Stanford.

DROUET Isabelle, 2012, *Causes, probabilités, inférences*, Paris, Vuibert, 154 p.

DUHEM Pierre, 1906, *La théorie physique. Son objet et sa structure*, Paris, Chevalier & Rivière.

DUPRÉ John, 2002, *Human Nature and the Limits of Science*, Oxford, Oxford University Press, 218 p.

FAN Jianqing, HAN Fang et LIU Han, 2014, « Challenges of Big Data analysis », *National Science Review*, 1 juin 2014, vol. 1, nº 2, p. 293-314.

FAURE-FREMIET E, 2017, « Les origines de l'Académie des sciences de Paris (1840-1945) », *Quaderns d'Història de l'Enginyeria*, 2017, vol. 15, p. 20-31.

FAYYAD Usama, PIATETSKY-SHAPIRO Gregory et SMYTH Padhraic, 1996, « From data mining to knowledge discovery in databases », *AI magazine*, 1996, vol. 17, nº 3, p. 37-54.

FEW Stephen, 2007, « Save the pies for dessert », *Visual Business Intelligence Newsletter*, août 2007 p.

FIELDING N. G. et LEE R. M., 1997, « Applications of Computer Software in the Sociological Analysis of Qualitative Data », *Bulletin de Méthodologie Sociologique*, 1 décembre 1997, vol. 57, nº 1, p. 3-24.

FOUETILLOU Guilhem, 2007, « Le web et le traité constitutionnel européen », *Réseaux*, 2007, vol. 144, nº 5.

FRANKLIN Allan, 1994, « How to avoid the experimenters' regress », *Studies in History and Philosophy of Science*, 1994, vol. 25, nº 3, p. 463-491.

FRAWLEY William J, PIATETSKY-SHAPIRO Gregory et MATHEUS Christopher J, 1992, « Knowledge Discovery in Databases : An Overview », *AI Magazine*, 1992, vol. 13, nº 3, p. 57-70.

FRIPPIAT Didier et MARQUIS Nicolas, 2010, « Les enquêtes par Internet en sciences sociales : un état des lieux », *Population*, 2010, vol. 65, nº 2, p. 309-338.

FUKUSHIMA Kunihiko, 1980, « Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position », *Biological Cybernetics*, 1980, vol. 36, nº 4, p. 193-202.

GAYO-AVELLO Daniel, 2012, « "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data », 28 avril 2012.

GELMAN Andrew et SHALIZI Cosma Rohilla, 2013, « Philosophy and the practice of Bayesian statistics. », *The British journal of mathematical and statistical psychology*, février 2013, vol. 66, nº 1, p. 8-38.

GEORGES Fanny, 2009, « Représentation de soi et identité numérique », *Réseaux*, 2009, vol. 154, nº 2, p. 165.

GIGERENZER Gerd Gerg, SWIJTINK Zeno, PORTER Theodore, DASTON Lorraine, BEATTY John et KRÜGER Lorenz, 1989, *The Empire of Chance. How Probability Changed Science and Everyday Life*, New York, Cambridge University Press.

GINSBERG Jeremy, MOHEBBI Matthew H, PATEL Rajan S, BRAMMER Lynnette, SMOLINSKI Mark S et

BRILLIANT Larry, 2009, « Detecting influenza epidemics using search engine query data. », *Nature*, 2009, vol. 457, nº 7232, p. 1012-1014.

GINZBURG Carlo, 1980, « Signes, traces, pistes », *Le Débat*, 1980, nº 6, p. 3-44.

GOLUMBIA David, 2009, *The cultural logic of computation*, Cambridge, Massachusetts, Harvard University Press.

GOMEZ-URIBE Carlos A et HUNT Neil, 2015, « The Netflix Recommender System : Algorithms, Business Value, and Innovation », *ACM Transactions on Management Information Systems*, 2015, vol. 6, nº 4, p. 1-19.

GOODY Jack, 1979, *La raison graphique : la domestication de la pensée sauvage*, Paris, Editions de Minuit, 274 p.

GRAHAM Mark, 2013, « Geography / Internet: Ethereal Alternate Dimensions of Cyberspace or Grounded Augmented Realities? », 2013, p. 1-14.

GREFENSTETTE Gregory, 1994, « Corpus-derived First, Second, and Third-order Word Affinities », Meylan, France.

GROLEMUND Garrett et WICKHAM Hadley, 2014, « A Cognitive Interpretation of Data Analysis », *International Journal of Statistics*, 2014, vol. 82, nº 2, p. 184-204.

GROSSBERG Stephen, 1988, « Nonlinear neural networks: Principles, mechanisms, and architectures », *Neural Networks*, 1988, vol. 1, nº 1, p. 17-61.

GUERIN-PACE France, 1997, « La statistique textuelle. Un outil exploratoire en sciences sociales », *Population*, 1997, vol. 52, nº 4, p. 865-887.

GUICHARD Éric, 2013, « L'internet et les épistémologies des SHS », *Sciences/Lettres*, 2013, nº 2.

HACKING Ian, 2002, *L'émergence de la probabilité*, Paris, Editions du Seuil.

HACKING Ian, 1999, *The Social Construction of What?*, Cambridge, Massachusetts, Harvard University Press.

HACKING Ian, 1981, « Est-ce qu'on voit à travers un microscope ? » dans Sandra Laugier et Pierre Wagner (eds.), *Philosophie des sciences - Tome 2 : Naturalismes et réalismes*, New York, Vrin, p. 305-322.

HALEVY Alon, NORVIG Peter et PEREIRA Fernando, 2009, « The Unreasonable Effectiveness of Data », *IEEE Intelligent Systems*, 2009, vol. 24, nº 2, p. 8-12.

HEIDELBERGER Michael, 2003, « Theory-Ladenness and Scientific Instruments in Experimentation » dans Hans Radder (ed.), *The Philosophy of Scientific Experimentation*, Pittsburg, University of

Pittsburgh Press, p. 138-151.

HEY Tony, TANSLEY Stewart et TOLLE Kristin, 2009, *The Fourth Paradigm*, Redmond, Microsoft Research, 287 p.

HOOKWAY Christopher, 2013, « Pragmatism » dans Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2016 Edition)*, Palo Alto, Stanford University.

HORNIK Kurt, STINCHCOMBE Maxwell et WHITE Halbert, 1989, « Multilayer feedforward networks are universal approximators », *Neural Networks*, 1989, vol. 2, nº 5, p. 359-366.

HUMMON Norman P. et FARARO Thomas J., 1995, « The emergence of computational sociology », *The Journal of Mathematical Sociology*, 1995, vol. 20, nº 2-3, p. 79-87.

HUMPHREYS Paul, 2004, *Extending ourselves. Computational science, empiricism and scientific method*, Oxford, Oxford University Press.

ISRAEL-JOST Vincent, 2016, « Computer Image Processing: An Epistemological Aid in Scientific Investigation », *Perspectives on Science*, 2016, vol. 24, nº 6, p. 669-965.

JEBEILE Julie, 2016, « Les simulations sont-elles des expériences numériques? », *Dialogue-Canadian Philosophical Review*, 2016, vol. 55, nº 1, p. 59-86.

JEBEILE Julie, 2013, *Explication et compréhension dans les sciences empiriques*,Université de Paris 1 Panthéon-Sorbonne, Thèse de doctorat en philosophie.

JOUISON-LAFFITTE Estèle, 2009, « La recherche action : oubliée de la recherche dans le domaine de l'entrepreneuriat », *Revue de l'Entrepreneuriat*, 2009, vol. 8, nº 1, p. 1.

JURGENSON Nathan, 2012, « When Atoms Meet Bits: Social Media, the Mobile Web and Augmented Revolution », *Future Internet*, 2012, vol. 4, nº 1, p. 83-91.

KANT Emmanuel, 1995, *Critique de la faculté de juger*, Paris, Aubier.

KITCHIN R. et MCARDLE G., 2016, « What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets », *Big Data & Society*, 2016, vol. 3, nº 1, p. 1-10.

KITCHIN Rob, 2014, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, http://www.uk.sagepub.com/books/Book242780?subject=F00&fs=1 , 2014, consulté le 19 août 2014.

KOSINSKI Michal, STILLWELL David et GRAEPEL Thore, 2013, « Private traits and attributes are predictable from digital records of human behavior », *Proceedings of the …*, 9 avril 2013, vol. 110, nº 15, p. 2-5.

KRÄMER Sybille, 2012, « Qu'est-ce donc qu'une trace, et quelle est sa fonction épistémologique ? État des lieux », *Trivium. Revue franco-allemande de sciences humaines et sociales*, 2012, vol. 10, nº 2012.

KWAK Haewoon, LEE Changhyun, PARK Hosung et MOON Sue, 2010, « What is Twitter, a social network or a news media? », New York, ACM Press.

LAGOZE Carl, 2014, « Big Data, data integrity, and the fracturing of the control zone », *Big Data & Society*, 13 novembre 2014, vol. 1, nº 2.

LALLICH-BOIDIN Geneviève, 2001, « Données linguistiques et traitement des questions ouvertes », *Journal de la Société française de statistique*, 2001, vol. 4, p. 29-36.

LANCIANI Albino Attilio, 2011, *Analyse phénoménologique du concept de probabilité*, Paris, Hermann.

LANDAUER Thomas K et DUMAIS Susan T, 1997, « A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge », *Psychological Review*, 1997, vol. 1, nº 2, p. 211-240.

LANEY Doug, 2001, *3D data management: controlling data volume, variety and velocity (META Group File 949)*, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data- Volume-Velocity-and-Variety.pdf , 2001.

LASSEGUE Jean, 1996, « La méthode expérimentale, la modélisation informatique et l'intelligence artificielle », *Intellectica*, 1996, vol. 22, nº 1, p. 21-65.

LAZER David, KENNEDY Ryan, KING Gary et VESPIGNANI Alessandro, 2014, « The Parable of Google Flu: Traps in Big Data Analysis », *Science*, 14 mars 2014, vol. 343, nº 6176, p. 1203-1205.

LEBART L et SALEM A, 1994, *Statistique textuelle*, Paris, Dunod.

LECUN Yann, 2014, *Deep learning. Week 1: Intro to Deep Learning*, http://cilvr.cs.nyu.edu/lib/exe/fetch.php?media=deeplearning:dl-intro.pdf , 2014.

LEONELLI Sabina, 2016, *Data-Centric Biology: A Philosophical Study*, Chicago, University of Chicago Press.

LEONELLI Sabina, 2014, « What difference does quantity make? On the epistemology of Big Data in biology », *Big Data & Society*, 1 avril 2014, vol. 1, nº 1, p. 2053951714534395.

LEWIN K, 1952, *Field Theory in Social Science*, Londres, Tavistock.

LEYDESDORFF Loet, 1995, *The challenge of scientometrics. The development, measurement and self-organization of scientific communications*, Leiden University, DSWO Press.

LOEVE Sacha, 2009, *Le concept de technologie à l'échelle des molécules-machines*,Université de Paris-

Ouest, Thèse de doctorat en philosophie, 682 p.

LOOSVELDT Geert et SONCK Nathalie, 2008, « An evaluation of the weighting procedures for an online access panel survey », *Survey Research Methods*, 2008, vol. 2, nº 2, p. 93-105.

MACQUEEN James, 1967, « Some methods for classification and analysis of multivariate observations », *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, vol. 1, nº 233, p. 281-297.

MÄKI Uskali, 2013, « Scientific Imperialism: Difficulties in Definition, Identification, and Assessment », *International Studies in the Philosophy of Science*, 2013, vol. 27, nº 3, p. 325-339.

MANOVICH Lev, 2011, « Trending: The promises and the challenges of big social data », *Debates in the Digital Humanities*, 2011, p. 1-17.

MANOVICH Lev, 2007, *Cultural analytics: analysis and visualization of large cultural data sets*, http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Cultural+analytics:#2 , 2007, consulté le 24 décembre 2012.

MARC X, 2001, « Les modalités de recueil des réponses libres en institut de sondage : le rôle de l'enquêteur, les consignes et les procédures de contrôle, les perspectives d'amélioration », *Journal de la Société française de statistique*, 2001, vol. 142, nº 4, p. 21-28.

MARR David, 1982, *Vision : a computational investigation into the human representation and processing of visual information*, Cambridge, Massachusetts, The MIT Press, 397 p.

MARRES Noortje, 2012, « The redistribution of methods: on intervention in digital social research, broadly conceived », *The Sociological Review*, 2012, vol. 60, p. 139-165.

MAYER-SCHÖNBERGER Viktor et CUKIER Kenneth, 2013, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, New York, Eamon Dolan/Houghton Mifflin Harcourt.

MAYO Deborah G. et COX David, 2009, « New Perspectives on (Some Old) Problems of Frequentist Statistics », *Error and Inference*, 2009, vol. 49, p. 247-330.

MCCOSKER Anthony et WILKEN Rowan, 2014, « Rethinking 'big data' as visual knowledge: the sublime and the diagrammatic in data visualisation », *Visual Studies*, 2014, vol. 29, nº 2, p. 155-164.

MCCULLOCH Warren S. et PITTS Walter, 1943, « A logical calculus of the ideas immanent in nervous activity », *The Bulletin of Mathematical Biophysics*, 1943, vol. 5, nº 4, p. 115-133.

MERZEAU Louise, 2013a, « L'intelligence des traces », *Intellectica*, 2013, vol. 59, nº 1, p. 1-22.

MERZEAU Louise, 2013b, « Traces numériques et recrutement : du symptôme au cheminement » dans Béatrice Galinon-Mélénec et Sami Zlitni (eds.), *Traces numériques : de la production à l'interprétation*,

Paris, CNRS éditions, p. 35-53.

METAXAS PT, 2012, « Why Is the Shape of the Web a Bowtie? », Lyon.

METROPOLIS Nicholas et ULAM S., 1949, « The Monte Carlo Method », *Journal of the American Statistical Association*, septembre 1949, vol. 44, nº 247, p. 335-341.

MIKOLOV Tomas, CHEN Kai, CORRADO Greg et DEAN Jeffrey, 2013, « Distributed Representations of Words and Phrases and their Compositionality » dans C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani et K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, New York, Curran Associates.

MONNIN Alexandre, 2012, « Les ressources, ces ombres récalcitrantes. », *SociologieS [En ligne], Théories et recherches*, 2012.

MORETTI Franco, 2000, « Conjectures on World Literature », *New Left Review*, 2000, nº 1, p. 54-68.

MORSTATTER Fred, PFEFFER J, LIU Huan et CARLEY KM, 2013, « Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose », *Proceedings of ICWSM*, 2013.

MOUNIER Pierre, 2015, « Une « utopie politique » pour les humanités numériques ? », *Socio*, 25 avril 2015, nº 4, p. 97-112.

MUSSO Pierre, 2010, « Le Web : nouveau territoire et vieux concepts », *Réalités industrielles*, 2010.

MUSSOU Claude, 2012, « Et le Web devint archive : enjeux et d{é}fis », *Le Temps des m{é}dias*, 2012, nº 19, nº 2, p. 259-266.

O'HARA Kieron et HALL Wendy, 2012, « Web Science » dans William Dutton (ed.), *The Oxford Handbook of Internet Studies*, Oxford, Oxford University Press.

PAK Alexander et PAROUBEK Patrick, 2010, « Twitter as a Corpus for Sentiment Analysis and Opinion Mining », *LREC*, 2010, p. 1320-1326.

PEGNY Maël, 2013, *Sur les limites empiriques du calcul. Calculabilité, complexité et physique*,Université Paris 1 Panthéon-Sorbonne, Thèse de doctorat en philosophie des sciences.

PEGNY Maël, 2012, « Les deux formes de la thèse de Church-Turing et l'épistémologie du calcul », *Philosophia Scientiæ*, 2012, vol. 1, nº 3, p. 39-67.

PEIRCE Charles Sanders, 1878, « How to Make Our Ideas Clear », *Popular Science Monthly*, 1878, vol. 12, January, p. 286-302.

PERINI Laura, 2012, « Truth-bearers or Truth-makers? », *Spontaneous Generations: A Journal for the History and Philosophy of Science*, 2012, vol. 6, nº 1, p. 142-147.

PETIT Victor, 2017, « Perspectives sur le design. Métier, enseignement, recherche », *Cahiers COSTECH*, 2017, vol. 1.

PETIT Victor et DELDICQUE Timothée, 2017, « La recherche en design avant la « recherche en design » », *Cahiers COSTECH*, 2017, vol. 1, nº 1.

PINCEMIN B, 2012, « Sémantique interprétative et textométrie », *Texto! Textes et Cultures*, 2012, vol. 17, nº 3, p. 1-21.

POPPER Karl Raimund, 1985, *Conjectures et réfutations : la croissance du savoir scientifique*, Paris, Payot, 610 p.

PORTER Theodore M., 2003, « Statistics and Statistical Methods » dans Theodore M. Porter et Dorothy Ross (eds.), *The Cambridge History of Science*, Cambridge, Cambridge University Press, vol.7, p. 238-250.

PORTER Theodore M., 1995, *Trust in Numbers*, Princeton, Princeton University Press, 325 p.

PROVOST Foster et FAWCETT Tom, 2013, « Data Science and its Relationship to Big Data and Data Driven Decision Making », *Big Data*, 2013, vol. 1, nº 1, p. 51-59.

RAJARAMAN Anand, ULLMAN Jeffrey David JD, LESKOVEC Jure, RAJARAMAN Anand et ULLMAN Jeffrey David JD, 2011, *Mining of Massive Datasets*, Cambridge, Cambridge University Press.

RASTIER François, 2015, « Sémantique de corpus — Questions d'épistémologie et de méthodologie. », *Texto! Textes et Cultures*, 2015, vol. 20, nº 1, p. 1-11.

RASTIER François, 2001, *Arts et sciences du texte*, Paris, Presses Universitaires de France.

RATINAUD Pierre et DEJEAN Sébastien, 2009, « IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre », Toulouse.

REICHFELD Fred et MARKEY Rob, 2011, *The Ultimate Question 2.0 (Revised and Expanded Edition): How Net Promoter Companies Thrive in a Customer-Driven World*, Cambridge, Massachusetts, Harvard Business Review Press.

REINERT M., 1990, « Alceste une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval », *Bulletin de Méthodologie Sociologique*, 1990, vol. 26, nº 1, p. 24-54.

RICKERT Heinrich, 1997, *Science de la culture et science de la nature suivi de Théorie de la définition*, Paris, Gallimard.

RICOEUR Paul, 1983, *Temps et récit 1*, Paris, Seuil.

RIEDER Gernot et SIMON Judith, 2017, « Big Data: A New Empiricism and its Epistemic and Socio-

Political Consequences » dans *Berechenbarkeit der Welt?*, Wiesbaden, Springer Fachmedien Wiesbaden, p. 85-105.

RIES Eric, 2011, *The lean startup : how today's entrepreneurs use continuous innovation to create radically successful businesses*, New South Wales, Currency, 320 p.

ROBINSON Viviane M J, 1993, « Current controversies in action research », *Public Administration Quarterly*, 1993, vol. 17, nº 3, p. 263-290.

RODRIGUES Maria De Lurdes, 2017, « La profession d'ingénieur au Portugal », *Quaderns d'Història de l'Enginyeria*, 2017, vol. 15, p. 271-286.

ROGERS Richard, 2013, *Digital methods*, Cambridge, Massachusetts, The MIT Press.

ROGERS Richard, 2010, « Internet Research: The Question of Method—A Keynote Address from the YouTube and the 2008 Election Cycle in the United States Conference », *Journal of Information Technology & Politics*, 2010, vol. 7, nº 2-3, p. 241-260.

ROGERS Richard, 2009, « The end of the virtual: Digital methods », Amsterdam.

ROGERS Richard, 2008, « Post-Demographic Machines », 2008.

ROSCH Eleanor H., 1973, « Natural categories », *Cognitive Psychology*, 1 mai 1973, vol. 4, nº 3, p. 328-350.

ROSENBLATT F, 1958, « The perceptron: a probabilistic model for information storage and organization in the brain. », *Psychological review*, novembre 1958, vol. 65, nº 6, p. 386-408.

RUMELHART David E., HINTON Geoffrey E. et WILLIAMS Ronald J, 1985, « Learning internal representations by error propagation » dans David E. Rumelhart et J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundation*, Cambridge, Massachusetts, MIT Press.

SAHLGREN M, 2008, « The distributional hypothesis », *Italian Journal of Linguistics*, 2008, vol. 20, nº 1, p. 33–54.

SAVAGE M. et BURROWS R., 2007, « The Coming Crisis of Empirical Sociology », *Sociology*, 2007, vol. 41, nº 5, p. 885-899.

SCHINDLER Samuel, 2013, « Theory-laden experimentation », *Studies in History and Philosophy of Science Part A*, 2013, vol. 44, nº 1, p. 89.

SCHLICK Moritz, 2003, *Formes et contenus*, Marseille, Agone.

SCHMITT Eglantine, 2015, « La structuration disciplinaire et thématique des humanités numériques »,

Strasbourg.

SCHONLAU Matthias, ZAPERT Kinga, SIMON Lisa P, SANSTAD Katherine Haynes, MARCUS Sue M, ADAMS John, SPRANCA Mark, KAN Hongjun, TURNER Rachel et BERRY Sandra H, 2004, « A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey », *Social Science Computer Review*, 1 février 2004, vol. 22, nº 1, p. 128-138.

SCHURMANS Marie-Noëlle, 2011, *Expliquer, interpréter, comprendre: le paysage épistémologique des sciences sociales*, Genève, Carnets des sciences de l'éducation.

SCHUTT Rachel et O'NEIL Cathy, 2014, *Doing Data Science*, , nº 9, Sebastopol, Californie, O'Reilly, vol.53.

SERRES A, 2002, « Quelle (s) problématique (s) de la trace? », Saint-Etienne.

SHADBOLT Nigel, HALL Wendy, HENDLER James A et DUTTON William H, 2013, « Web science: a new frontier », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 28 mars 2013, vol. 371, nº 1987.

SHINN Terry et RAGOUET Pascal, 2000, « Formes de division du travail scientifique et convergence intellectuelle . La recherche technico-instrumentale », *Revue française de sociologie*, 2000, vol. 41, nº 3, p. 447-473.

SILVER Nate, 2012, *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't*, Londres, Penguin, 544 p.

SIMONDON Gilbert, 2005, *L'invention dans les techniques. Cours et conférences*, Paris, Seuil.

SIMONDON Gilbert, 1958, *Du mode d'existence des objets techniques*, Paris, Aubier.

SNOW C. P., 1959, « Two Cultures », *Science*, 21 août 1959, vol. 130, nº 3373, p. 419-419.

SPANOS Aris et MAYO Deborah G., 2015, « Error statistical modeling and inference: Where methodology meets ontology », *Synthese*, 2015.

STEVENS Kent A., 1981, « The visual interpretation of surface contours », *Artificial Intelligence*, août 1981, vol. 17, nº 1-3, p. 47-73.

STIGLER Stephen M., 1999, *Statistics on the Table: The History of Statistical Concepts and Methods*, Cambridge, Massachusetts, Harvard University Press.

STRASSER Bruno J, 2012, « Data-driven sciences: From wonder cabinets to electronic databases. », *Studies in history and philosophy of biological and biomedical sciences*, 2012, vol. 43, nº 1, p. 85-87.

SU Yu-Sung, 2008, « It's easy to produce chartjunk using Microsoft®Excel 2007 but hard to make good

graphs », *Computational Statistics & Data Analysis*, 2008, vol. 52, nº 10, p. 4594-4601.

TANGUY Ludovic, 2013, « La ruée linguistique vers le Web », *Texto !*, 2013, vol. 18, nº 4, p. 1-33.

TAYLOR L., SCHROEDER R. et MEYER E., 2014, « Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? », *Big Data & Society*, 2014, vol. 1, nº 2.

THIERRY B, 2013, *Donner à voir, permettre d'agir. L'invention de l'interactivité graphique et du concept d'utilisateur en informatique et en télécommunications en France (1961-1990)*,Paris-Sorbonne (Paris IV), Thèse de doctorat en histoire contemporaine.

TIERCELIN Claudine, 1993, *C.S. Peirce et le Pragmatisme*, Paris, Presses Universitaires de France, 128 p.

TUFTE Edward R., 2008, *The Visual Display of Quantitative Information*, Cheshire, Connecticut, Graphics Press, 1-191 p.

TUFTE Edward R., 1997, *Visual explanations*, Cheshire, Connecticut, Graphics Press.

TUKEY John W., 1977, *Exploratory Data Analysis*, Reading, Massachussets, Addison-Wesley, 506 p.

TURING Alan M. A.M., 1940, « Psychology and Philosophy », *Nature*, 27 avril 1940, vol. 145, nº 3678, p. 662-662.

TUROW Joseph, MCGUIGAN Lee et MARIS Elena R., 2015, « Making data mining a natural part of life : Physical retailing , customer surveillance and the 21st century social imaginary », *European Journal of Cultural Studies*, 2015, vol. 18, nº 4-5, p. 464-478.

VANDENDORPE Christian, 1999, *Du papyrus à l'hypertexte*, Paris, La Découverte.

VAPNIK Vladimir Naumovich, 1998, *Statistical learning theory*, New York, Wiley, 736 p.

VARENNE Franck, 2007, *Du modèle à la simulation informatique*, Paris, Vrin.

VARENNE Franck et SILBERSTEIN Marc, 2013, *Epistémologies et pratiques de la modélisation et de la simulation*, Paris, Editions Matériologiques, 982 p.

VAYRE Jean-Sébastien, 2014, « Manipuler les données. Documenter le marché », *Les Cahiers du numérique*, 2014, vol. 1, p. 1-30.

VIAL Stéphane, 2016, « Le tournant design des humanités numériques », *Revue française des sciences de l'information et de la communication*, 2016.

WANG Yilun et KOSINSKI Michal, 2017, « Deep neural networks are more accurate than humans at detecting sexual orientation from facial images », *Journal of Personality and Social Psychology (JPSP)*, 2017, p. 1-47.

WIGNER Eugene P., 1960, « The unreasonable effectiveness of mathematics in the natural sciences »,

Communications on Pure and Applied Mathematics, février 1960, vol. 13, nº 1, p. 1-14.

WINDELBAND Wilhelm, 2000, « Histoire et sciences de la nature. Discours prononcé au Rectorat de Strasbourg », *Les Études philosophiques*, 2000, vol. 1.

WING Jeannette M, 2006, « Computational thinking », *Communications of the ACM*, 2006, vol. 49, nº 3, p. 33.

WITTE J. C., 2009, « Introduction to the Special Issue on Web Surveys », *Sociological Methods & Research*, 1 février 2009, vol. 37, nº 3, p. 283-290.

WITTGENSTEIN Ludwig, 2004, *Recherches philosophiques*, Paris, Gallimard.

WU Xiaolin et ZHANG Xi, 2016, « Responses to Critiques on Machine Learning of Criminality Perceptions (Addendum of arXiv:1611.04135) », 2016.

WU Xindong, KUMAR Vipin, ROSS QUINLAN J., GHOSH Joydeep, YANG Qiang, MOTODA Hiroshi, MCLACHLAN Geoffrey J., NG Angus, LIU Bing, YU Philip S., ZHOU Zhi-Hua, STEINBACH Michael, HAND David J. et STEINBERG Dan, 2008, « Top 10 algorithms in data mining », *Knowledge and Information Systems*, 4 janvier 2008, vol. 14, nº 1, p. 1-37.

YAKOUT Mohamed, BERTI-ÉQUILLE Laure et ELMAGARMID Ahmed K., 2013, « Don't be SCAREd: Use SCalable Automatic REpairing with Maximal Likelihood and Bounded Changes », *Proceedings of the 2013 international conference on Management of data - SIGMOD '13*, 2013, nº 4, p. 553.

YEUNG Karen, 2017, « Algorithmic Regulation : A Critical Interrogation », *Regulation and Governance*, 2017, p. 1-39.

ZHAO Yanchang, 2012, « R and Data Mining: Examples and Case Studies », 2012.