> Eglantine Schmitt

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

(Doctorat)

> #Numéro 2
> Doctorats, Habilitations
> Thèses de doctorat
> EPIN - Ecritures, Pratiques et Interactions Numériques (Costech-UTC)
> Mégadonnées - > Méthodologie de recherche - > Philosophie

Références de citation

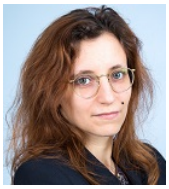
Schmitt, Eglantine. "Explorer, visualiser, décider : un paradigme méthodologique pour la production de connaissances à partir des big data. (Doctorat)", 14 janvier 2019, màj 0000, *Cahiers COSTECH*http://www.costech.utc.fr/CahiersCOSTECH/spip.php?article73

Thèse de doctorat en Epistémologie et histoire des sciences et des techniques, soutenue à l'Université de Technologie de Compiègne, le 14 novembre 2018.

Résumé

L'enjeu de cette thèse de philosophie des sciences était de répondre à un problème pratique, qui s'est présenté à Églantine alors qu'elle travaillait comme analyste sur des données massives chez Proxem : comment produire des connaissances valides en manipulant de grandes masses de données qu'elle n'a pas constituées et qui ne sont pas le fruit d'une méthode scientifique reconnue ?

En s'appuyant sur son expérience de terrain, Églantine 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, elle nous 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.



Eglantine Schmitt est Docteure en Épistémologie et histoire des sciences et des techniques. Elle est aujourd'hui Head of Product chez prevision.io, éditeur de solutions d'intelligence prédictive pour l'entreprise. En s'appuyant notamment sur ses recherches en sciences des données, son rôle est de démocratiser les technologies d'intelligence artificielle en concevant des outils prédictifs accessibles, simples d'utilisation, et qui répondent aux problèmes concrets des utilisateurs métier.

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. Il 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.

+

Le fichier de la thèse en texte intégral ici :

Doctorat E.Schmitt, 2018 (texte intégral, accès ouvert)
Thèse de doctorat en Epistémologie et histoire des sciences et des techniques, soutenue à l'Université de Technologie de Compiègne, le 14 novembre 2018.