Le numérique comme milieu : enjeux épistémologiques et phénoménologiques

Principes pour une science des données

< Bruno Bachimont >

Sorbonne universités, Université de Technologie de Compiègne, CNRS
Laboratoire Heudiasyc, centre de recherche Royallieu, 60205 Compiègne cedex
bruno.bachimont@utc.fr

< RESUME >

Les mégadonnées constituent une mutation majeure dans notre rapport de représentation scientifique du monde. Nous argumentons que nous passons d'une épistémologie de la mesure, sur laquelle la science de la nature est fondée, à une épistémologie de la donnée, permettant d'aborder à nouveaux frais le monde de la culture. Ce passage introduit la nécessité d'aborder deux chantiers : l'un épistémologique, l'autre phénoménologique. Le premier exige de comprendre en quoi les représentations construites permettent de mettre en place un régime de connaissance et pas seulement de représentation. Le second correspond au fait d'articuler l'expliquer au comprendre dans le cadre des traitements effectués sur des mégadonnées rassemblées autour d'activités et faits humains.

< ABSTRACT >

Big data are a major revolution in the field of science where an epistemology of data follows an epistemology of measure. This epistemological revolution concerns as a first place social and human sciences insofar as human facts and activities are represented by data on which mathematical treatments are applied. Two major issues emerge. One concerns epistemology because one should understand how big data propose new knowledge and new discoveries ; a second concerns phenomenology insofar as big data should preserve the meaning of social and human facts even through the mediation of mathematical treatments.

< MOTS-CLES >

Epistémologie des données, mégadonnées.

< KEYWORDS >
Epistemology of data, big data

1. Introduction

Le terme de « numérique » est souvent utilisé non seulement comme un substantif mais aussi comme un adjectif introduisant l'ambiguïté de savoir s'il doit fonctionner comme un génitif objectif ou subjectif ; apporte-t-il une précision sur ce qu'il qualifie dans sa nature (génitif subjectif) ou son objet (génitif objectif) : une linguistique numérique par exemple, parle-t-elle du numérique, ou utilise-t-elle les outils numériques pour aborder ses objets habituels ? Une philologie numérique aborde-t-elle les problèmes philologiques posés par le numérique ou veut-elle traiter ses objets habituels en enrichissant ses méthodes d'outils numériques ?

La mobilisation du numérique s'effectuant la plupart du temps depuis un point de vue instrumental, le numérique semble souvent n'être qu'un outil, dont l'utilisation massive ne remet pas en cause son statut instrumental ni la manière dont il faudrait le considérer. Le génitif subjectif prime.

Pourtant, le numérique n'est pas seulement un nouveau moyen d'aborder des problèmes anciens, mais d'envisager des problèmes inédits, de poser des problèmes qui n'existeraient pas dans un contexte où cette puissance technique serait absente. Cette caractéristique, le numérique la partage avec la plupart des techniques, qui par nature contribuent à inventer des possibles et à poser des questions qui seraient impensables et insensées sinon. Mais ce qui rend le numérique singulier, c'est qu'il le fait à un niveau de généralité, d'universalité et d'homogénéité rarement rencontré sinon pour des inventions techniques exemplaires comme l'écriture.

En effet, le numérique est à la fois une technologie intellectuelle et une ingénierie pour les systèmes physiques. Par technologie intellectuelle, à la suite de Levy (2000) et Stiegler (1994), on désigne l'ensemble des procédés et outils permettant d'instrumenter l'expression, la communication, l'interprétation et plus généralement ce

qui relève du monde du symbole et du signe. Par ingénierie pour les systèmes physiques, on évoque les techniques permettant de construire les machines et de transformer la matière sous ses différentes formes. Comme en témoigne l'*annus mirabilis* de 1943 où deux articles fondateurs ont mis l'information et le calcul à la base des systèmes physiques (Wiener *et al.,* 1943) d'une part et des cerveaux pensants d'autre part (McCulloch et Pitts, 1943), le numérique unifie des sphères techniques jusque là séparées ou aux liens distendus : le monde de la transformation physique de la matière, de l'énergie et du mouvement, et le monde de la manipulation du signe et des formes culturelles. De l'internet des objets qui permet de mettre un calculateur au sein de tout artefact physique aux systèmes de représentation qui reposent désormais tous sur le calcul et la manipulation des signes, le numérique n'est plus seulement un outil permettant de réaliser nos fins dans le monde, mais le milieu (Beaune, 1998) à travers lequel nous y accédons et la médiation pour le penser et y agir.

Il ne s'agit donc pas de savoir comment se servir du numérique, mais de comprendre le monde que nous construisent nos outils numériques, outils qui sont autant pour la pensée que pour l'action. Comment pensons nous, comment agissons-nous dans le monde, comment pouvons nous le connaître et y agir quand le milieu qui nous permet de nous y rattacher, articuler, opposer ou fusionner est désormais de nature numérique ?

On peut distinguer deux manières complémentaires de considérer le numérique : d'une part, comme l'introduction du calcul dans notre manière de présenter rationnellement le monde, d'autre part comme l'introduction d'outils et de représentations fonctionnant comme des interfaces à travers lesquels nous nous rapportons au monde. La première approche, plus épistémologique, aborde la question de savoir ce que devient la science quand ses principes scientifiques sont désormais gouvernés par ceux du calcul au lieu des principes classiques telles la mathématisation dans les sciences de la nature ou l'interprétation ou la critique dans les sciences de la culture. La seconde est davantage phénoménologique dans la mesure où il s'agit de comprendre en quoi le numérique manifeste des configurations de sens nouvelles et redéfinit les modalités d'appréhension sous lesquelles nous

abordons des objets de penser, de perception ou d'imagination inédits. Appelons la première l'approche computationnelle, dont l'enjeu est le principe d'une raison computationnelle, la seconde l'approche numérique à proprement parler : en effet, nous avons introduit le calcul depuis longtemps dans nos pratiques scientifiques et culturelles sans qu'on parle de rupture ; c'est depuis que l'on numérise nos contenus culturels, nos outils, nos processus de manière intégrée que nous sommes désormais dans la situation d'aborder n'importe quel objet en mobilisant du calcul sans nous en rendre compte, le calcul étant désormais caché dans les artefacts que nous manipulons. Pour le dire autrement, le numérique apparaît quand d'autres choses que les calculatrices et les ordinateurs (dont la finalité est le calcul) mobilisent le numérique pour d'autres finalités que le calcul.

Ces deux dimensions sont également présentes dans ce qu'on appelle désormais de plus en plus une science des données. Elles sont importantes pour appréhender tant ce qui constitue la scientificité de cette nouvelle approche que pour comprendre les perspectives de son utilisation.

L'objectif de cet article est d'aborder ces deux aspects pour déterminer les principaux enjeux, promesses ou menaces d'une telle science.

2. De la mesure à la donnée : une mutation épistémologique

2.1. Le paradigme classique : de la mesure à la théorie

La révolution scientifique moderne, dont les principaux héros traditionnels sont Galilée, Descartes et Newton, repose sur le fait de mobiliser l'expérimentation instrumentée et le calcul pour construire une nouvelle appréhension de la nature qui permette d'en faire une représentation rationnelle soumise à des lois. L'édification de la science moderne repose donc sur la mesure et son enrôlement dans des lois mathématiques et calculatoires permettant d'en déduire une compréhension et prédiction des phénomènes. On est ainsi dans le contexte de ce que nous appelons une « épistémologie de la mesure », où sont précisées les conditions et les principes sous lesquels est justifié le

fait que la connaissance scientifique se construit de manière fiable et raisonnée à partir de la mesure.

Plusieurs gestes philosophiques ont contribué à comprendre ces conditions, de Bacon (2001) à Bachelard (2000) en passant par Kant (1980). C'est ce dernier que nous allons mobiliser ici pour comprendre et illustrer ce qui fait qu'une science de la mesure fonctionne. Comme on le sait, Kant, fasciné par les réussites éclatantes de la science newtonienne, ne cherche pas à construire une science de la nature, celle de Newton existe, mais à comprendre pourquoi et comment une telle science est possible. Son approche, certes célèbre pour sa complexité, peut être – sans doute exagérément – simplifiée pour notre propos.

En effet, selon Kant, la nature se manifeste à nous à travers ce que l'esprit appréhende sous la forme d'un « divers phénoménal » ou un divers espace-temps :

> *Dans le phénomène, je nomme matière de celui-ci ce qui correspond à la sensation, tandis que ce qui fait que le divers du phénomène peut être ordonné selon certains rapports, je le nomme la forme du phénomène.* (AK, IV, 30), (Kant 1997).

Le phénomène se manifeste pour l'esprit comme une diversité donnée et non produite par le sujet, diversité qu'il faut dès lors rapporter à une unité, ce qui fait qu'on y reconnaît ou comprend quelque chose (le fait de voir une maison par exemple). Cette construction de l'unité se fait selon un cadre *a priori*, l'espace et le temps, dont la structure renvoie à l'infini et au continu mathématiques. Révélé comme une quantité dans un cadre mathématique, le divers kantien est en fait une mesure. Par la suite, Kant explique comment ce divers se rapporte aux concepts de la raison grâce à des schémas qui sont des formes spatiotemporelles découlant du contenu scientifique des concepts. Autrement dit, les concepts scientifiques se schématisent sous la forme de lois mathématiques qui viennent s'appliquer sur les mesures (le divers).

Ce petit rappel kantien (voir par exemple (Ferry, 2008) et (Rivelaygue, 1992) pour des présentations précises et lumineuses) nous suffit ici. Car ce que nous voulons illustrer en l'occurrence, c'est le fait

élémentaire et simple qu'il y a une continuité, une homogénéité réunissant dans une même cohérence l'information révélée par la nature (le divers ou la mesure effectuée par l'artefact expérimental) et la loi qui s'applique sur ce divers pour y voir la réalisation de tel concept ou telle loi de la physique par exemple. Ou, dit autrement, les hypothèses et les théories permettant de constituer les appareils de mesure et les outils d'interprétation sont les mêmes dans les deux cas. Dans le vocabulaire bachelardien, la phénoménotechnique, qui est le fait que les phénomènes scientifiques sont donnés par des appareils de mesure reposant pour leur conception et réalisation sur des lois scientifiques, appartient à la même cohérence que l'interprétation scientifique des phénomènes : ce sont les mêmes lois qui permettent de construire les appareils de mesure et d'interpréter ce qu'ils délivrent.

2.2. Un autre paradigme : de la donnée à la visualisation

Si on s'intéresse à présent à ce que l'on appelle la science des données (ce qu'on peut voir thématisé par exemple dans (Hey *et al.*, 2009), on constate que les principes de fonctionnement de l'épistémologie de la mesure ne sont pas réunis ici. En effet, si on reprend les principes de fonctionnement des approches fondées sur les données, comme Lev Manovich les énonce par exemple (Manovich, 1996) en voulant présenter de nouvelles disciplines comme les *Cultural Analytics,* on trouve trois étapes fondamentales :

– la collecte des données, qui repose sur le fait de capter des informations d'origines diverses, selon des périodicités élevées, et hétérogènes dans leur nature et dans leur format ; on retrouve le principe des 4 « V » des big data : Velocity, Variability, Volume et Verity ;

– le traitement des données, qui repose sur l'utilisation d'outils en général mathématiques et statistiques ;

– la visualisation des résultats, qui repose sur la présentation des résultats selon des conventions souvent cartographiques, s'inspirant des techniques de domaine de l'InfoViz (*Information Visualisation*).

Ces trois étapes, aussi simplement formulées, permettent déjà cependant de tirer des conclusions importantes. D'une part, il n'existe pas d'hypothèse particulière sur la nature des données captées, en

particulier elles ne constituent pas un divers au sens kantien, c'est-à-dire une manifestation s'exprimant dans un cadre a priori, de nature mathématique, qui lui donne sa forme. En effet, les données ne sont pas des mesures : elles sont des informations, c'est-à-dire des expressions formatées selon une structure formelle ou symbolique de manière arbitraire, et collectées grâce à leur format. On peut donc avoir des logs de consultation d'un site web, des textes collectés, des transactions, etc., alors qu'elles n'appartiennent pas au même cadre *a priori*. On prend donc ici la notion d'in*format*ion dans le sens précis du contenu formaté, dans la mesure où le formatage permet la manipulation formelle qui rend possible leur captation et leur traitement. La donnée est alors la plus petite partie manipulable de l'information comme contenu formaté.

Deuxièmement, le traitement des données s'applique selon des principes qui sont encore une fois totalement arbitraires par rapport à la nature des données. Dans la reconstruction que fait Kant par exemple de la science classique, les mesures permettent de construire l'objet scientifique selon les lois et concepts scientifiques car la mesure et la loi appartiennent au même cadre, s'articulent de manière complémentaire dans une même conception et construction de la science. Or, quelle que soit la constitution des données, leur nature, leur collecte, dès l'instant que l'on sait les organiser dans un tableau, on pourra leur appliquer des algorithmes cherchant à expliquer leur distribution, faire ressortir les cas fréquents, isolés, etc. La loi scientifique n'appartient pas à la même construction scientifique que la donnée.

Enfin, la visualisation des données et des résultats mobilise des modes de présentation dont l'interprétation est différente dans ses principes de ce qui a permis de collecter les données d'une part, et de les traiter d'autre part. Peu importe ce que l'on nous montre, on regarde une organisation esthétique et sémiotique de données (par exemple cartographique) où les principes de compréhension, la manière de donner du sens à la carte ou la représentation, sont arbitraires par rapport aux données et à leur traitement. Cependant, cette organisation est véritablement productive au niveau du sens, au niveau de la réception et perception des formes (esthétiques) et du registre sémiotique adopté (stéréotypes culturels notamment).

Cela implique que les sciences des données sont dans une triple rupture : une rupture des données par rapport à leur origine et leur nature, le mode de collecte les rassemblant en dépit de leur hétérogénéité ; une rupture du traitement par rapport aux données ; une rupture de ce qui est montré par rapport à ce qui est calculé.

Ces ruptures sont à la fois ce qui constitue la force des sciences des données, et ce qui en est la faiblesse. Puisqu'il y a rupture entre les données et leur origine, on peut les collecter de manière arbitraire par rapport à celle-ci et réunir ce qui est différent et devrait donc le rester selon les critères classiques. On peut faire des rapprochements inédits, bouleverser les interdits ou les clivages disciplinaires ou sociotechniques, pour réunir ce qui, sinon, est séparé. De même, la rupture au niveau de la présentation esthétique des résultats est ce qui permet de surmonter la complexité et la masse des traitements, qui sont difficilement maîtrisables dans leur formalisation mathématique, et inintelligibles dans leur mise en œuvre étant donné la masse de calculs et de donnés mobilisés. Sans une telle rupture, on serait noyé et perdu. Enfin, la rupture au niveau des traitements est également essentielle : parce qu'on est agnostique sur les données, on peut les traiter quelle que soit leur origine.

Indiscutablement, ces ruptures sont potentiellement des progrès majeurs et essentiels. Elles s'inscrivent dans l'histoire des outils et technologies de synthèse des inscriptions de la pensée : par exemple, l'écriture a permis la synopsis de la parole en l'inscrivant sur un support matériel effectuant la syn-thèse, c'est-à-dire le fait de poser (thèse) ensemble (syn) et simultanément les composants dispersés dans la succession temporelle de la parole (Bachimont, 2010). Les progrès techniques permettent de proposer au regard de l'esprit des éléments qui seraient sinon dispersés et d'y reconnaître des configurations de sens indécelables sans ces médiations techniques. Les techniques numériques, les « big data », reposent sur ce principe : en rapprochant l'hétérogène, ce dont l'origine, la conformation technique contribuaient à séparer et à maintenir dans des univers différents, se retrouve désormais juxtaposé et manipulé dans un même espace pour une synthèse à la fois spatiale (le même espace de stockage) et calculatoire (la mobilisation dans de mêmes calculs).

Mais on découvre de ce fait que la scientificité de la science des données ne peut reposer sur le modèle traditionnel, puisqu'il n'y a pas de cadre commun entre les données, les lois qui les mobilisent et l'esthétique de leur représentation ; rien ne permet de comprendre en quoi le résultat fourni est scientifique, et encore moins quel statut lui donner et quelle interprétation en faire.

Par conséquent, la science des données n'est pas un nouvel outil pour aborder des questions anciennes permettant de résoudre par de nouveaux moyens des problèmes difficiles par leur taille ou leur masse. Par ces ruptures, la science des données est une rupture épistémologique : il est donc impératif de construire l'épistémologie à sa hauteur, ce que nous appelons une épistémologie de la donnée.

2.3. Construire une épistémologie de la donnée

La science des données oppose à la *motivation* du paradigme classique, qui fonde la mise en œuvre de la mesure dans une interrogation du réel qui découle des lois et principes qui permettront de l'interpréter, *l'arbitraire* qui pose la donnée, la manipule et enfin présente une visualisation, opérations qui ne sont plus motivées par la nature du réel et les hypothèses qu'on a formulées à son égard.

Cet arbitraire est nécessaire comme on l'a dit : rapprocher des réalités hétérogènes (collecte), surmonter la masse des données (traitements mathématiques) et la complexité des calculs effectués (visualisation et esthétisation des résultats). Mais que cet arbitraire soit justifié ne suffit pas à en définir la nature. Or, cet arbitraire est rendu possible par la définition même du calcul et de l'information comme substrat binaire.

Le numérique se construit autour d'une ascèse du signe, où le symbole manipulé par le calcul ne se définit plus que par sa seule manipulabilité, indépendamment de toute réalité qui lui serait associée. Le numérique se caractérise en effet par une double coupure :

– Une coupure matérielle, selon laquelle le symbole numérique est indépendant de la manière dont il est réalisé matériellement, autrement

dit de son implémentation ; cette propriété est ce qui nous permet de croire que nous avons le même fichier sur des substrats physiques différents, comme un disque dur ou un DVD de sauvegarde ; en effet, alors que ces fichiers sont physiquement différents (l'un correspond à des trous et des bosses d'un support optique, l'autre à des tensions magnétiques fortes ou faibles), ils sont numériquement identiques car ces reliefs (trous/bosses) et tensions codent la même chose, des 0 et des 1.

– Une coupure sémantique, selon laquelle le symbole numérique ne prescrit aucune interprétation particulière : un même flux binaire peut être lu comme une image ou comme un son par exemple.

Le symbole numérique est donc un pur donné, un fait, un absolu qui n'a aucun lien avec ce qui pourrait l'avoir produit, avec ce qui pourrait le concrétiser et enfin l'interpréter. Le paradoxe du terme de « donné » est que la donnée ne permet pas de savoir de quoi elle est la donnée et par qui elle est donnée. Autonome vis-à-vis d'un qui et d'un quoi, la donnée est le pur arbitraire, le pur fait, une priméité percienne (Peirce, 1978).

Cette double coupure a pour conséquence que les corrélations établies entre les données viennent combler le manque de relation avec l'origine d'où la donnée est issue et le monde réel où l'interprétation est censée la replonger. La réalité est prise en charge par la corrélation calculée entre les données, sans qu'il soit possible de contrôler leur véracité : on est conduit à apprécier la plausibilité de ces liens à défaut de pouvoir rapporter la pertinence du calcul à des mesures se confrontant à la réalité du monde. Le calcul effectué sur les données est par conséquent incommensurable au monde, au sens étymologique du terme, au sens où il n'y pas de commune mesure entre la donnée et le monde dont elle est issue ; ce qu'elle dit du monde est un arbitraire établi par la corrélation sans qu'il soit possible de confronter le contenu sémantique attribué à la donnée aux conditions de son extraction, sa construction et son traitement.

Il est frappant que le monde nouveau des données massives ne nous ait pas encore appris quelque chose que nous ignorions : on est bien plutôt dans un régime de confirmation statistique d'interprétation dont on avait déjà la notion. S'agit-il tout simplement d'un manque de

maturité de ces nouvelles techniques, ou plutôt d'un positionnement fautif dans la chaîne interprétative reliant le monde et son interprétation ? La question n'est pas tranchée. Mais il est patent que le traitement de la donnée reste encore indéterminé et partagé entre le régime de la confirmation – justification ou de la découverte, selon la célèbre distinction établie en son temps par Hans Reichenbach (1938).

3. De la donnée à l'interprétation : une mutation phénoménologique

La discussion précédente a permis de mettre en évidence que la donnée se construit comme une rupture et un arbitraire incommensurables avec le monde dont elle est pourtant issue dans sa construction et auquel elle doit être reconduite par son interprétation. La question est alors de comprendre comment se négocie cet arbitraire et quel est le régime du sens qui s'y construit.

La principale séduction des mégadonnées vient de leur capacité d'intégrer malgré leur hétérogénéité de multiples informations se rapportant à l'activité humaine. L'arbitraire de la donnée qui constitue un défaut d'origine (la donnée dans sa nature numérique est coupée de son origine causale et factuelle) lui donne un supplément d'interprétabilité en lui permettant d'être associée et confrontée à des informations différentes.

L'enjeu et la révolution attendue des mégadonnées sont donc bien la compréhension nouvelle promise des activités humaines. Les mégadonnées pourraient constituer une révolution des sciences de la culture à l'instar de la révolution scientifique qui a permis, à l'orée de notre modernité, de transformer notre relation à la nature d'une description fondée sur le langage à un rapport de mesure expérimentale et de formalisation calculée. À l'instar de la nominalisation des sciences de la nature, les mégadonnées proposent la nominalisation des sciences de la culture.

Le nominalisme est cette école de pensée, dont on connaît les héros/hérauts principaux à travers des figures comme Guillaume d'Ockham, Jean Buridan, Jean Gerson, etc., qui consiste à considérer que

les noms généraux de catégories comme « humanité », « animalité », ne renvoient à aucune entité existante (Libéra, 1993). Il n'existe que des entités singulières, les noms généraux comme humanité n'étant alors que des *signes* permettant de se référer à ces dernières de manière collective, mais sans avoir pour autant d'entité associée comme une humanité qui serait une partie ontologique de chaque être humain. Le nominalisme apporte donc une solution *sémiotique* et non plus *ontologique* au problème de traiter les termes généraux (Libéra, 1994).

Le nominalisme s'oppose au réalisme qui considère que de telles entités existent, non pas de manière séparée ou autonome, mais au sein des individus qu'elles désignent et constituent (Erismann, 2011). Le réalisme repose *in fine* sur l'idée que les structures fondamentales du langage renvoient à la structure du réel, la logique du discours étant, correctement étudiée, une logique du réel (Paqué, 1985). La langue est un accès au réel et son intelligibilité est une intelligence de la nature : la structure syntaxique du langage renvoie à la structure ontologique du réel. Quand on dit par exemple « l'homme est un animal », cela signifie que l'essence « humanité » contient en sa définition l'essence de l'« animalité ».

La révolution nominaliste a rompu le lien ontologique entre la langue et la nature, laissant ainsi la place pour qu'une nouvelle relation s'établisse entre notre concept et la connaissance de la nature : cette relation sera la mesure expérimentale exprimée dans l'idiome mathématique, permettant ainsi à la légalisation formelle d'exprimer la structure ontologique du monde. Ce que Galilée saura exprimer en disant dans l'*Essayeur* en 1623 (Galilée 1980) que :

> *La philosophie est écrite dans ce grand livre qui se tient constamment ouvert devant les yeux (je veux dire l'Univers), mais elle ne peut se saisir si tout d'abord on ne se saisit point de la langue et si on ignore les caractères dans lesquels elle est écrite. Cette philosophie, elle est écrite en langue mathématique ; ses caractères sont des triangles, des cercles et autres figures géométriques, sans le moyen desquels il est impossible de saisir humainement quelque parole ; et sans lesquels on ne fait qu'errer vainement dans un labyrinthe obscur.*

Les mégadonnées proposent une semblable mutation concernant le monde de la culture : au lieu d'interpréter le fait culturel par la compréhension qu'on en a à travers l'expression qu'il reçoit dans le langage, il s'agit désormais de calculer des corrélations entre les données rassemblées. Alors que la nominalisation de la nature remplaçait la description linguistique par la mesure expérimentale et le calcul, le nominalisme de la culture remplace la compréhension linguistique par la donnée collectée et la corrélation statistique.

Or, la révolution scientifique a non seulement transformé notre rapport à la nature, mais la notion même de nature, en la désenchantant comme l'a si lucidement constaté Max Weber (2002). Il en est probablement de même de notre notion de culture et notre rapport à l'activité humaine et au fait humain.

En effet, étudier un fait humain[1], qu'est-ce que cela signifie ? Comme le formule suggestivement Marc Bloch dans son *apologie pour l'histoire* (Bloch, 1997), le scientifique de la culture traque le fait humain comme l'ogre des histoires enfantines : il flaire l'humanité, là est son gibier. Or l'humanité, c'est le fait de reconnaître que le fait qui arrive est un fait qui arrive à un alter ego, c'est un fait qui aurait pu m'arriver et qui prend sens pour moi dans l'altérité de la distance culturelle et temporelle. L'historien et l'anthropologue, pour prendre ces deux figures des scientifiques qui étudient respectivement la distance temporelle et la distance spatio-culturelle, pratiquent cette herméneutique particulière d'une construction théorique de l'Autre à partir d'un investissement à partir de soi et du présent (Marrou, 1954). Autrement dit, l'intelligibilité du fait culturel repose sur l'empathie qui permet de construire la mise à distance à partir d'une assimilation à ce qui aurait pu m'arriver, et en comprenant que justement, ce ne m'est pas arrivé et ne pourra pas l'être : que ce soit le contexte qui diffère, ma manière de penser qui l'interdit, les limites du pensable et du faisable, l'assimilation problématique au présent et à soi est le moyen privilégié de construire l'altérité et la mise à distance. Comme le dit joliment Antoine Prost, expliquer le terroir médiéval à une personne qui n'a jamais connu à la première personne ce qu'est un terroir est une entreprise sans espoir

1. Nous reprenons ici une réflexion que nous avons entamée dans (Bachimont, 2014).

puisque la meilleure manière de comprendre de quoi on parle est de se rapporter à cette expérience vécue pour comprendre et saisir que ce qu'on vise est quelque chose de bien différent (Prost, 1996).

L'investigation du fait humain ne se réduit pas à l'empathie bien sûr, mais la complète par une investigation empirique et logique du fait humain comme fait, en le confrontant par l'analyse à son contexte empirique et à la cohérence qu'il entretient avec ce que l'on sait déjà. Des séries statistiques à l'indice permettant de remonter une série de causes ou d'analyse, le scientifique de la culture pratique lui aussi le recours aux sciences formelles pour interroger la réalité qu'il étudie. Mais toutes les statistiques sur le cours du pain à la veille de la révolution ne permettront en rien d'expliquer à quiconque n'a jamais eu faim ni ne sait ce que c'est que manger du pain frais ou cuit au four, en quoi le cours du pain peut avoir une quelconque relation avec la révolution, les tensions sociales, les émeutes populaires, bref la misère qui fait descendre dans la rue. On pourrait renvoyer ce dernier aspect au vécu, au ressenti, au subjectif, ce qui n'est alors pas de l'ordre de la connaissance. C'est notamment la thèse du Cercle de Vienne et de son principal héraut, Moritz Sclick (1985), ce qui lui permet en particulier de considérer qu'on peut avoir un cadre unique pour la science, unissant les sciences de la nature et les sciences de l'esprit. Mais justement, la question est qu'en faisant cela, on perd l'intelligibilité du lien qui existe entre les faits mobilisés, lien reposant sur le fait humain sous jacent. En cela, les big data sont très proches du Cercle de Vienne, qui ne recherche pas les relations de causalité dans le monde réel, mais des corrélations formelles (logiques, mathématiques, statistiques) : de même que les sciences contemporaines ignorent la causalité pour comprendre le monde de la nature, les big data conduisent à ignorer le fait humain pour comprendre le monde de la culture et ce qui relie les êtres humains. On pense ainsi gagner en précision, mais a-t-on davantage expliqué ? C'est certainement là une question qui sera encore débattue, entre la puissance d'agir fondée sur la corrélation précise mais aveugle, et l'intelligibilité qui permet de comprendre sans forcément permettre l'efficacité décisionnelle ou prédictive (Thom, 2009).

On comprend alors pourquoi les études sur les *big data* en font trop ou pas assez : elles vont trop loin car elles perdent l'humanité des faits

qu'elles étudient. Si bien que les résultats brandis sont bien des propriétés des modèles d'analyse, mais ne nous apprennent que fort peu sur la réalité dont les données sont issues. Et donc elles ne vont pas assez loin : pour tenir le programme qu'elles se sont donné, il leur faudrait aller au bout de l'interprétation et comprendre en quoi les analyses menées permettent de retrouver l'humanité des faits étudiés et ainsi avancer dans la connaissance et la compréhension de la culture.

4. Conclusion : pour une science des données et ses enjeux épistémologiques et phénoménologiques

La science des données est une mutation majeure et une promesse d'un regard nouveau et d'une intelligibilité inédite sur le monde des faits humains. Mais il convient de prendre cette mutation au sérieux. Selon les considérations précédentes, deux chantiers sont à envisager.

Le premier est un chantier épistémologique : que connaît-on à travers les mégadonnées ? Quel monde se montre à travers les interprétations effectuées au terme des visualisations ? Quelles sont les modalités de vérification, justification des faits établis par ces traitements ? S'agit-il de confirmer statistiquement ce que l'on sait déjà ou de découvrir ce qu'on ignore ? Mais les mégadonnées sont ramenées au célèbre paradoxe du Ménon de Platon : comment peut-on en effet apprendre quelque chose de nouveau ? Si quelque chose de nouveau se présente, comment le reconnaître ? Et si je le reconnais, c'est qu'il n'est pas nouveau. Quand, en face des cartes déduites des traitements effectués, on construit des interprétations, comment savoir ce que l'on voit sinon en le rapportant à ce qu'on sait déjà ? Puisque le lien au monde est dans l'opacité de la complexité de la collecte et du traitement, les seules transparence et interprétabilité possibles viennent de ce que nous savons déjà reconnaître et projeter en guise de plausibilité. Les mégadonnées doivent être travaillées pour devenir des instruments de connaissance et pas seulement de reconnaissance ou de justification.

Le second chantier est phénoménologique : comment retrouver dans les traitements effectués sur les données l'humanité des faits examinés ? En quoi les activités examinées sont-elles des activités humaines ? On comprend bien la différence entre l'histoire du cosmos qu'on établit en

physique et l'histoire des faits humains qu'élabore la science historique. Les mégadonnées ne risquent-elles pas de nous faire passer de la seconde à la première et ainsi de perdre la promesse qu'elles nous annoncent ?

Les réponses à cette question ne sont pas encore élaborées, mais des orientations déjà suggestives. En particulier, on a pris l'habitude d'articuler les pratiques de *close reading* et celles du *distant reading* dans l'étude de grands corpus de textes (Moretti, 2008) : alors que la première repose sur une lecture interprétative faisant droit à la compréhension linguistique et l'empathie qu'elle recèle, la seconde permet de thématiser les effets de la globalité d'un corpus, d'observer à travers une représentation calculée des effets de sens particuliers. Au lieu de les opposer, il s'agit de la composer : les allers et retours entre ces deux postures de lecture permettant d'accéder aux apports du traitement des données et de retrouver la compréhension d'une donnée singulière en son sein.

En conclusion, il s'agit d'apprendre à *lire* ces mégadonnées selon de nouvelles pratiques à dégager : de même que Ben Schneidermann avait popularisé dans l'univers des interfaces numériques le slogan « global and zoom » (Bederman and Schneidermann, 2003), un axe d'étude des mégadonnées est de construire une composition lectoriale du global et du local : la singularité herméneutique abordée au sein de la globalité calculée et restituée dans des visualisations graphiques.

Bibliographie

Bachimont Bruno (2014). Le nominalisme et la culture : questions posées par les enjeux du numérique. In *Digital studies : Organologie des savoirs et technologies de la connaissance,* Bernard Stiegler (ed). Paris, FYP Editions.

Bachimont Bruno (2010). *Le sens de la technique : le numérique et le calcul*. Paris, Les Belles Lettres.

Bachelard Gaston (2000). *La formation de l'esprit scientifique*. Paris, Vrin.

Bacon Francis (2001). *Novum Organum.* Paris, PUF.

Beaune Jean-Claude (1998). *Philosophie des milieux techniques : La matière, l'instrument, l'automate*. Seyssel, Champ Vallon.

Bederson Benjamin B., Schneiderman Ben (Eds.). (2003). *The Craft of Information Visualization: Readings and Reflections*. Morgan Kaufmann.

Bloch Marc (1997). *Apologie pour l'histoire ou Métier d'historien.* Paris, Armand Colin.

Erismann Christian (2011). *L'homme commun : la genèse du réalisme ontologique durant le haut Moyen-Âge*. Paris, Vrin.

Ferry Luc (2008). *Kant : une lecture des trois « Critiques ».* Paris, Le livre de poche.

Galilei Galileo (1980). *Il Saggiatore*, traduction française de Christiane Chauviré, *L'Essayeur*, Paris, Les Belles-Lettres.

Hey Tony, Tansley Stewart, Tolle Kristin (2009). The Fourth Paradigm : Data-Intensive Scientific Discovery. Redmond, Microsoft Research.

Kant Emmanuel (1997). *Critique de la raison pure* (Traduction Alain Renaut). Paris, Aubier.

Lévy Pierre (1990). *Les technologies de l'intelligence. L'avenir de la pensée à l'ère informatique.* Paris, La Découverte.

Libera Alain de (1994). *La querelle des universaux : de Platon à la fin du Moyen Âge.* Paris, Seuil.

Libera Alain de (1993). *La philosophie médiévale.* Presses Universitaires de France.

Manovich Lev (2008). *Cultural analytics: analysis and visualisation of large cultural data sets.* Software Studies Initiative.

Marrou Henri-Irénée (1954). *De la connaissance historique.* Paris, Seuil.

McCulloch Warren and Pitts Walter (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics,* Oxford, Elsevier Sciences, 5, 115-133.

Moretti Franco (2008). *Graphes, cartes et arbres : modèles abstraits pour une autre histoire de la littérature.* Paris, Les Prairies ordinaires.

Paqué Ruprecht. (1985). *Le Statut parisien des Nominalistes.* Paris, Presses Universitaires de France.

Peirce Charles Sanders (1978). *Écrits sur le signe.* Paris, Seuil.

Prost Antoine (1996). *Douze leçons sur l'histoire.* Paris, Seuil.

Reichenbach Hans (1938). *Experience and Prediction*. Chicago, University of Chicago Press.

Rivelaygue Jacques (1992). *Leçons de métaphysique allemande, Tome II : Kant, Heidegger, Habermas.* Paris, Grasset.

Rosenbluth Arturo., Wiener Norbert., and Bigelow Julian. (1943). Behavior, purpose and Teleology. *Philosophy of Science,* Baltimore, Williams & Wilkins, vol. X, p. 18-24.

Schlick Moritz. (1985). Le vécu, la connaissance, la métaphysique. In A. Soulez (Ed.), *Manifeste du Cercle de Vienne et autres écrits* (pp. 183-197). Paris, Presses Universitaires de France.

Stiegler Bernard (1994). *La technique et le temps ; Tome I : la faute d'Epiméthée.* Paris, Galilée.

Thom René (2009). *Prédire n'est pas expliquer*. Paris, Champ.

Weber Max (2002). *Le savant et le politique*. Paris, 10/18.