TRACES, CALCUL ET INTÉRPRETATION: DE LA MESURE À LA DONNÉE

1. *Introduction.*

Dans le processus qui nous a conduit à la révolution numérique que nous connaissons à présent, il est sans doute utile d'avoir un regard rétrospectif pour tenter de saisir les principales étapes et ruptures qui nous ont amenés là où nous en sommes à présent. Sans retracer bien évidemment l'histoire du numérique et du calcul, on peut déterminer des étapes clefs et indiquer quelques points de repère chronologiques et conceptuels.

1.1. *Du calcul aux contenus dématérialisés: les âges du numérique.*

Les principales étapes que nous distinguons sont au nombre de trois: le calcul comme objet mathématique, le calcul comme machine matérielle, le calcul comme médiation culturelle.

Au début du siècle dernier, Hilbert[1] et Turing[2] ont en effet contribué à donner une description mathématique de ce qu'est un calcul, et plus généralement, une démonstration, un théorème, c'est-à-dire les objets qui correspondent à l'activité du mathématicien et pas seulement à ses objets d'étude. Hilbert a pu parler de métamathématique dans la mesure où les mathématiques se prenaient elles-mêmes comme objet, ses méthodes devenant ses objets. Mais si le calcul recevait enfin une définition formelle et précise (ce

[1] D. Hilbert, *Sur les fondements de la logique et de l'arithmétique*, trad. Hourya Sinaceur, dans *Logique et fondements des mathématiques*, Paris, Payot, 1992, pp. 255-275.

[2] A. M. Turing, *Théorie des nombres calculables, suivi d'une application au problème de la décision*, in *La machine de Turing*, sous la direction de J.-Y. Girard, Paris, Seuil, 1995, pp. 49-104.

ESL053_Azimuth_7.indd 13 06/06/16 12:15

que fait une machine de Turing), il restait à matérialiser cette conception théorique en une possibilité technique et mécanique.

La possibilité de principe fut donnée par l'article séminal de McCulloch et Pitts en 1943[3], *annus mirabilis* qui vit la naissance conceptuelle de deux courants de pensée qui allait marquer notre modernité, la cybernétique avec l'article de Wiener, Biegelow et Rosenbluth[4] sur la téléologie et la rétroaction d'une part, les neurones formels et la mécanisation de la pensée avec McCulloch d'autre part. En effet cet article ne disait rien moins que le cerveau, objet biologique matériel, pouvait être décrit de manière idéalisée, via les neurones formels, et exhiber ainsi les mêmes possibilités qu'une machine de Turing. Le concept théorique de calcul pouvait donc rencontrer une matérialité naturelle (le cerveau) et artificielle (les nouvelles machines qui seraient appelées plus tard «ordinateur»).

La dernière étape fut donc le déploiement de calcul comme technologie au sein de nos industries et entreprises. Là encore, il convient de cerner plusieurs étapes, les deux premières ayant été déjà relevées par Allan Newell et Herbert Simon[5]:

– l'ordinateur comme calculateur (*number cruncher*); l'enjeu est alors le calcul scientifique;
– l'ordinateur comme processus symbolique (*symbolic processor)*; l'enjeu est alors l'information et sa gestion;
– l'ordinateur comme medium universel: l'enjeu est alors la manipulation de contenu et la gestion des interactions.

D'un point de vue global, ces trois étapes renvoient à des âges industriels que nous avons vu se succéder: l'informatique des scientifiques puis l'informatique de gestion et enfin la révolution numérique. Cette dernière se caractérise par le fait que le numérique a réduit le calcul à sa plus simple expression, à savoir la simple manipulation d'unités discrètes (d'au moins de deux types, et en pratique seulement deux) sans y voir de nombres ni de symboles particuliers.

[3] W. McCulloch – W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, «Bulletin of Mathematical Biophysics», V (1943), pp. 115-133.

[4] A. Rosenbluth, N. Wiener, J. Bigelow, *Behavior, purpose and Teleology*, «Philosophy of Science», X (1943) pp. 18-24.

[5] A. Newell – H. A. Simon, *Computer Science as Empirical inquiry: Symbols and Search*, in *Mind Design*, sous la direction de J. Haugeland, Cambridge, The MIT Press, 1981, pp. 35-66.

ESL053_Azimuth_7.indd 14 06/06/16 12:15

1.2. *Le numérique comme mutabilité universelle et héraclitéenne des contenus.*

Le numérique est devenu un équivalent binaire universel de tout type de contenus, permettant ainsi de traduire n'importe quel contenu en un substrat numérique, soumis à la manipulation arbitraire et donc à la recomposition elle aussi arbitraire. C'est ainsi qu'un son, numérisé, devient un substrat binaire qu'on peut, si le loisir nous en prend, décoder comme une image et le projeter sur un écran, transformant ainsi un son en une image *via* la *même* ressource binaire. Le résultat est certes improbable et surprenant, mais possible, puisque le numérique est devenu ce medium anonyme et aveugle, permettant de coder n'importe quel contenu, et de le décoder en n'importe quel autre contenu, souvent le même que celui d'origine mais de moins en moins souvent du fait des transformations qu'il devient possible d'effectuer lors de la relecture / décodage d'une ressource.

Le numérique renouvelle le mobilisme universel d'Héraclite selon lequel on ne se baigne jamais deux fois dans le même fleuve: le numérique dissociant la ressource binaire stockée de la vue que l'on publie à partir d'elle, permet d'inventer à chaque fois une nouvelle vue puisque le décodage peut varier selon des critères contextuels d'une fois à l'autre. Le Web est là pour nous le rappeler et nous le montrer: la page consultée dépend des informations contextuelles stockées dans notre navigateur, de ses paramètres propres, et enfin des ressources à partir desquelles construire la page. Ainsi, on ne voit jamais deux fois le même contenu, un contenu n'est jamais montré deux fois de la même manière. Principe de mutabilité universel, le numérique devient la source des variantes infinies des contenus puisque qu'ils sont à chaque fois réinventés pour être montrés.

Le numérique est aux contenus ce que l'argent est aux biens économiques: un équivalent universel permettant de tout traduire, échanger, sans introduire de contraintes autre que celles imposées par l'arbitraire des conventions de codage et la réalisation effective des calculs de leur mise en œuvre.

1.3. *Le numérique comme milieu.*

Le numérique est non seulement une technique, mais aussi un milieu. Le milieu est ce qui nous environne et nous entoure (le 'milieu ambiant'), mais également ce qui est au milieu, c'est-à-dire ce qui nous relie. Le numérique est à la fois ce qui nous permet de nous rapporter à autrui, mais aussi à notre environnement, la médiation à l'autre et l'ailleurs.

Ces dernières années, se déploie de manière plus visible et manifeste les conséquences de ce paradigme de «milieu numérique». Equivalent

universel, le numérique permet de rassembler en son sein des contenus et informations de toute origine, de toute nature, rapportant dans une même homogénéité binaire des informations hétérogènes d'origines diverses. Le numérique est devenu de ce fait un outil remarquable de synthèse au sens où il permet de poser ('-thèse') ensemble ('syn-') des éléments qui seraient sinon éclatés et jamais rencontrés ensemble: des textes et des vidéos, des données administratives et scientifiques, etc.

Le numérique vient donc compléter la chaîne des technologies intellectuelles[6] qui depuis l'émergence de l'humain, contribuent à synthétiser des représentations de manière à pouvoir voir autrement le monde à travers elles, et y déceler des rapports, des connexions sinon indécelables, voire inexistantes si on adopte une posture constructiviste selon laquelle les représentations sont des médiations permettant la co-construction du sujet cognitif et de son environnement.

Synthèse cognitive donc, le numérique ne serait rien de plus qu'une écriture enrichie, puisque l'écriture n'est, en un sens, qu'une technique permettant de synthétiser dans un même espace d'inscription ce que la parole disperse dans la succession temporelle ou ce que la pratique éparpille dans ses différents espaces d'intervention. Mais alors que l'écriture permet d'offrir, à travers ses synthèses, une synopsis nouvelle au regard du sujet cognitif, cette synopsis (voir – *opsis*, ensemble – *syn*) ne proposant qu'une juxtaposition dont le regard percevant doit voir la pertinence et abstraire la structure[7], le numérique offre un espace de synthèse au calcul et à la manipulation, intégrant dans de grandes masses de données rendues homogènes par leur numérisation, des informations et contenus qui sont alors manipulables et qu'il est possible d'abstraire non par le regard perceptif mais par la procédure calculatoire[8].

Le numérique permet donc de passer de la synthèse synoptique à la synthèse calculante. Cette dernière livre alors une synthèse calculée au regard, aboutissant bien à une synopsis pour le regard perceptif, mais qui n'est plus la simple conséquence de l'écriture comme juxtaposition de contenu, mais le résultat de procédures abstraites et calculatoires.

[6] P. Lévy, *Les technologies de l'intelligence. L'avenir de la pensée à l'ère informatique*, Paris, La Découverte, 1990; B. Stiegler, *La technique et le temps. Tome I: la faute d'Epiméthée*, Paris, Galilée, 1994.

[7] Cf. J. Goody, *La raison graphique, la domestication de la pensée sauvage*, Paris, Les Editions de Minuit, 1979.

[8] Cf. B. Bachimont, *Le sens de la technique: le numérique et le calcul*, Paris, Les Belles Lettres, 2010.

ESL053_Azimuth_7.indd 16 06/06/16 12:15

De la synthèse synoptique de l'écriture à la synopsis calculée comme résultat de la synthèse calculante du numérique, telle est la transition que nous traversons actuellement. Cette mutation est complexe car elle pose un problème profond, à la hauteur des promesses qu'elle ouvre, celui de la transparence cognitive de la synopsis. En effet, l'écriture rapporte dans un même espace des éléments dispersés qui possèdent chacun leur pertinence et sens au sein de leur contexte propre. L'écriture, par le geste autoritaire et arbitraire du scripteur, construit un nouvel ordre du sens. Ce sens n'est pas toujours transparent: le livre sur les listes de Umberto Eco[9] est là pour nous rappeler combien l'action synthétisante de l'écriture peut paraître insolite et improbable. Cependant, notre raison graphique a construit des ordres de représentation et des types d'écriture où les synthèses sont devenues par elles-mêmes des constructions de sens, c'est-à-dire qu'elles construisent une cohérence et un ordre nouveau rendant intelligibles et compréhensibles les éléments rassemblés dans cette nouvelle disposition. Mais outre l'arbitraire que nous avons appris à maîtriser dans l'édification de notre civilisation graphique, l'écriture possède une propriété fondamentale: les éléments rassemblés restent toujours présents pour le regard et intelligibles pour eux-mêmes même si leur coexistence dans l'ordre graphique paraît insolite et improbable. Cette transparence intelligible est perdue dans la synthèse calculante du numérique: les données numériques sont masquées dans le résultat du calcul; de même la synthèse calculatoire introduit un type de transformation qui altère non seulement l'espace de représentation (comme le fait l'écriture) mais la donnée elle-même qui disparaît dans la transformation qui l'exploite.

La synopsis calculée pose donc un problème d'intelligibilité. En tant que synopsis, le résultat du calcul présente la difficulté inhérente à toute écriture de l'arbitraire de représentation dans un même espace. Mais en tant que synopsis calculée, on introduit en plus la difficulté de la transformation arbitraire de la donnée, en perdant son intelligibilité propre, liée à son origine, pour l'intégrer dans une procédure calculatoire. Ainsi des octets représentant des données nombrées de gestion par exemple, peuvent être intégrés à des octets représentants des pixels: une procédure calculatoire peut les mobiliser sans les distinguer. Il devient alors impossible de comprendre l'intelligibilité du résultat et encore moins de la rapporter à celle des données mobilisées.

On comprend alors l'urgence de s'intéresser aux conditions sous lesquelles il est possible de construire une intelligibilité des synopsis calculées

[9] U. Eco, *Vertige de la liste*, Paris, Flammarion, 2009.

à partir des synthèses calculantes du numérique. C'est l'urgence soulevée par les nouvelles sciences des données qui recherchent pour leur part à proposer de nouvelles médiations par le calcul pour synthétiser les masses de données.

Notre réflexion dans cet article est de comprendre ce que peuvent être ces conditions d'intelligibilité et ainsi de proposer une manière de procéder à une critique de la raison numérique puisque, si les sciences des données sont désormais un fait, il convient de les interroger sur leur bon droit, et passer du *de facto* au *de jure.*

2. *Traces, mesures, données.*

Le numérique est donc une synthèse calculante produisant une synopsis calculée. La synthèse rapproche dans un même espace de calcul des éléments collectés de manière arbitraires, des 'données'. Par ce terme, on entend qu'on convient de ne pas s'interroger sur l'origine de la donnée, sur sa nature: elle est là, donnée par on ne sait pas qui ni comment. Même s'il n'en est rien, bien évidemment, la synthèse commence là où le processus d'obtention s'arrête avec les questions que l'on peut se poser sur la définition de la donnée.

2.1. *Traces.*

Cette amnésie de l'origine propre à la donnée fait que la donnée est en rupture avec le paradigme de la trace et de la mesure. La trace est par définition le résultat de ce qui est causé par le comportement d'une entité dans son environnement. La trace a donc toujours une origine, une histoire, et l'enjeu est de pouvoir repérer la trace comme telle et ensuite de remonter de la trace à son origine. Une trace laissée dans l'environnement n'est pas toujours découverte; c'est pourquoi on pourra distinguer entre les notions suivantes:

- une trace est un état de l'environnement laissée par le passage d'un être, le comportement d'une entité ou le déroulement d'un processus;
- la trace peut être volontaire, involontaire ou provoquée:
  - › une trace involontaire est une empreinte,
  - › une trace volontaire est un message,
  - › une trace provoquée est une mesure.
- la trace, quand elle est détectée et reconnue comme telle, c'est-à-dire comme renvoyant à une cause qui l'a produite, devient un indice (on

s'écarte ici de la terminologie proposée par Alain Mille[10], où il appelle trace ce que nous appelons indice, et empreinte ce que nous appelons trace).

La trace est donc ce qui est produit et causé dans l'environnement par une origine, l'indice est ce qui reconnu par un interprète, un enquêteur. Ce dernier reconnaît les messages qu'on lui a adressés, et détecte les empreintes qu'on a laissées à son insu.

La trace est donc une signature et n'est jamais anonyme, elle porte en elle ce dont elle est trace, de manière plus ou moins explicite d'ailleurs. Avec la tradition antique (Platon dans le *Sophiste* et le *Gorgias*, Aristote dans la *Rhétorique*) on peut distinguer entre trois types de relation:

– la trace pathognomonique, celle que Platon appelait le *tekmerion*: si cette femme a du lait, c'est qu'elle a enfanté; il n'y pas de fumée sans feu. Dans ce cas, on dispose d'un lien causal, univoque, permettant de remonter de l'effet à la cause.
– La trace vraisemblable: le lien causal n'est plus univoque; bien des causes peuvent conduire à l'effet constaté, à la trace trouvée. Les antiques parlent alors de *eikos*. La trace montre son origine sans la déterminer. Par exemple, en séméiologie médicale, les signes cliniques sont pour la plupart des *eikoi*: s'il a chaud, c'est qu'il a de la fièvre.
– Enfin, la trace arbitraire: le lien n'est plus causal, mais seulement conventionnel. La trace n'est donc le signe de ce dont elle est trace que par la médiation du code, de la convention et de la connaissance qu'on en a. Il s'agit alors de *semeion*: si c'est rouge, alors c'est interdit.

Même quand le lien est arbitraire, il y a néanmoins un lien selon lequel la trace est intentionnelle et renvoie à autre chose qu'elle-même.

2.2. *Mesures.*

La mesure est dans ce contexte un cas particulier de trace, où un processus provoque une modification de l'appareil de mesure qui renseigne alors sur l'état de l'environnement. La mesure s'inscrit dans un contexte fortement théorisé où un corpus de connaissances permet de faire de lien entre la mesure obtenue et l'état de l'environnement que l'on peut en déduire: les

[10] A. Mille, *Des traces à l'ère du* Web, «Intellectica», LIX (2013), 1, pp. 7-28.

sciences de la nature ont développé une méthodologie particulière permettant ainsi de faire le lien.

Ce corpus théorique introduit un lien particulier entre la cause et l'effet, la mesure et le phénomène. Ce sont les mêmes hypothèses et concepts qui permettent de rendre raison du phénomène et de sa mesure: ces derniers sont homogènes car ils s'inscrivent dans une même compréhension et perspective du monde qui les contient. Les théories qui permettent de définir les explications rendant compte de la mesure sont les mêmes que celles qui ont permis de construire l'appareil de mesure: comme l'a souligné Bachelard[11], les appareils sont des théories matérialisées; ils appartiennent au même continuum de connaissances que les objets dont ils traquent les traces.

2.3. *Données.*

Les données ne sont pas des traces. Bien sûr, on utilise parfois des traces pour constituer des données. Mais le fait de se donner des 'données' renvoie à deux principes essentiels:

– en reprenant comme données ce qui initialement étaient des traces, on abolit l'histoire et l'origine des traces pour ne les considérer que dans la littéralité symbolique qui les constitue comme données: l'enregistrement formaté selon les règles de la synthèse calculante.
  › Par exemple, quand on considère les mots d'un texte pour faire une étude statistique quelconque, on oublie volontaire ce qu'est le texte dans son origine, signification, intention pour ne retenir que les mots qu'il contient en tant que ce sont des suites de caractères. On perd ainsi, volontairement, le fait que certains de ces mots étaient des titres, donc importants pour la signification et la compréhension du texte, que d'autres étaient dans des notes accessoires, donc plutôt secondaires, etc.
– Dans cet oubli volontaire, on rend commensurable ce qui ne l'était pas du fait des origines diverses. La commensurabilité signifie qu'il existe une mesure commune permettant de comparer, de mesurer les unes vis-à-vis des autres les entités commensurables. Quelle est donc ici, dans le contexte, des données, la mesure commune fondant la commensurabilité de toutes les données? La mesure est le format syntaxique et logique utilisé. Toutes les informations sont rapportées une même littéralité sym-

[11] G. Bachelard, *Le nouvel esprit scientifique*, Paris, Gallimard, 2003.

ESL053_Azimuth_7.indd 20 06/06/16 12:15

bolique qui les rend homogènes et agrégeables dans un même espace de calcul.

› Si on reprend le même exemple, les mots d'un texte deviennent tous des suites de caractères qui, comme telles, peuvent comparées (longueur, fréquence des lettres, flexions, etc.).

La littéralité symbolique a cet immense avantage, mais par conséquent inconvénient également, de permettre de rendre comparable via son formalisme (c'est là l'avantage) ce qui ne l'est pas de point de vue de son origine ou de sa signification (et c'est là l'inconvénient). Jadis, comme nous apprîmes à compter dans notre petite enfance, nos bons maîtres et diligentes maîtresses nous enseignèrent qu'il ne fallait pas compter des poires et des pommes ensemble sinon l'addition ne signifie plus rien. Si ce rappel et cet avertissement étaient aussi importants et pertinents, c'est précisément que dans sa neutralité ontologique et sémantique, le nombre ne préjuge pas de ce qu'il compte, et il est toujours possible d'agréger n'importe quel nombre avec n'importe quel autre. La numération arithmétique, la littéralité symbolique sont donc de même nature et transforment les informations ainsi formatées en données manipulables à loisir, indépendant de leur unité (si on compte), de leur dimension (si on mesure), de leur origine (si on récolte des indices).

La donnée instaure donc un nouvel ordre dans notre rapport au réel, rassemblant dans l'homogénéité du format l'hétérogénéité de l'origine et de la cause. La question est alors de savoir quelles conclusions on peut tirer des manipulations effectuées sur des données, et les connaissances que l'on peut en tirer.

3. *Théorie, modèles, calcul.*

Les données s'inscrivent dans un nouvel ordre symbolique et la question est de savoir comment les relations entre les données permettent de déduire des conséquences et connaissances sur ce dont les données sont tout de même les données, c'est-à-dire les traces, dans la multiplicité de leurs origines diverses. La manipulation des données correspond au calcul, les relations entre les résultats calculés et l'environnement sont données par la théorie qui encadre les calculs effectués et les modèles qui permettent de les utiliser. Il convient de préciser ce que l'on peut entendre par ces trois termes de théorie, modèle et calcul, pour comprendre en quoi le paradigme de la donnée, en congédiant l'instance théorique et reléguant la modélisation à un statut purement instrumental, nous propose quelque chose de radicalement nouveau même si il adopte les apparences habituelles du travail scientifique et technique. Ces trois notions peuvent se définir de la manière suivante:

– la théorie correspond au travail de compréhension du réel: la théorie a pour objectif de rendre intelligible son objet qu'elle contribue de ce fait à constituer. En s'intéressant à une sphère du réel ou des idées, la théorie propose des concepts qui rendent raison de cette sphère; la théorie est dès lors raisonnante et arraisonnante; elle rapporte le réel à ses concepts tout en permettant de l'appréhender et de s'y repérer.
  › Une théorie du vivant (sphère du réel) postulera par exemple que tout est génome, ce dernier étant le concept permettant de rendre intelligible cette réalité et par là même de la réduire à ce que la théorie peut en dire à travers ce concept;
– Le modèle est ce qui permet de se donner une représentation du réel pour utiliser cette représentation en lieu et place de la réalité représentée; lieutenant du réel, le modèle permet d'en faire abstraction car il en possède tous les attributs nécessaires. L'instar des ministres plénipotentiaires qui ont les pouvoirs de la puissance politique qu'il représente, le modèle peut répondre aux questions qu'on lui pose comme le ferait la réalité dont il assume et reprend les caractéristiques 'essentielles'.
  › Le modèle n'a pas de vocation universaliste ni globalisante: il ne prétend pas être une théorie du réel ou d'une sphère particulière de ce dernier; le modèle représente une situation donnée ou type de situation, comme par exemple un dispositif. On aura ainsi une théorie de la propulsion, mais le modèle d'un turbo-réacteur.
– Le calcul est ce qui permet de définir une manipulation formelle au sein d'une représentation formatée. Le calcul obéit à des procédures aveugles, des algorithmes, qui prescrivent des règles machinales qui peuvent dès lors être exécutées par une machine (!). La plupart du temps, le calcul applique ces algorithmes dans un cadre fixé par une théorie ou un modèle, déterminant ainsi le format des entités manipulées, les limites d'application des procédures, la nature des résultats obtenus. Le calcul ne s'applique pas sur le monde ou la réalité, mais toujours sur la représentation de ce dernier. C'est pourquoi le calcul appartient au domaine de la théorie et de la modélisation, car il permet de déployer la représentation théorique ou modélisatrice dans sa systématicité.

On a donc la théorie pour comprendre le réel en le rendant intelligible, le modèle pour agir par procuration comme si on avait le monde sous la main, le calcul pour manipuler le modèle ou la théorie.

Le modèle est une notion qui a pris une importance considérable au cours du XXe siècle dans la mesure où, outil privilégié de l'ingénieur, le modèle a trouvé dans les nouveaux outils numériques une manière de

déployer l'exploration systématique de ce qu'il affirme à propos du réel en prétendant le remplacer. On a de ce fait plusieurs types de modèles:

– des modèles qui permettent de décider:
  › La représentation doit permettre d'adopter une décision parmi des choix possibles. C'est là l'antique vocation du calcul qui, selon Aristote dans l'Ethique à Nicomaque (livre VI sur les vertus intellectuelles) [1965] permet de pondérer les arguments dans une délibération nécessaire car on n'est pas dans le domaine de la science ou de la démonstration («là où il y a démonstration, il n'est inutile de délibérer»). Ces modèles sont les outils privilégiés du gestionnaire et du manager.
– des modèles qui permettent de simuler:
  › Outil privilégié de l'ingénieur, le modèle pour simuler s'appuie sur une description physique du comportement étudié pour ensuite t simuler ce dernier en exécutant les calculs associés et prédire les évolutions possibles.
– des modèles qui permettent d'explorer:
  › Ce dernier type de modèle est proche de la théorie: dans ce cas le modèle reproduit suffisamment le réel pour permettre de l'explorer à travers lui, et gagner ainsi une meilleur compréhension de ce dernier. Le modèle peut renvoyer aux expériences de pensées (*Gedankenexperiment*) où la systématicité de la théorie postulée permet d'explorer les conséquences d'une situation imaginée.

A chaque fois, on comprend que la démarche habituelle articule les trois notions ensemble sans qu'aucune ne soit totalement autonome vis-à-vis des autres:

– la théorie fixe le cadre global, détermine les concepts, construit les objets et donne les lois et relations articulant ces derniers.
– Le modèle utilise la théorie pour décrire le comportement d'une situation particulière du monde réel (ou culturel).
– Le calcul met en œuvre le modèle et parcourt les inférences et déductions permettant de reproduire le comportement de la situation décrite par le modèle, que ce soit pour décider, simuler ou explorer.

Ainsi, le calcul n'a pas de sens sans le modèle qui le mobilise, le modèle sans la théorie qui lui fournit les principes de sa description. De même, une théorie sans modèle décrit le monde réel en général, mais ne permet pas

d'aborder les situations particulières, et un modèle sans calcul reste sans portée pratique.

4. *Le paradigme de la donnée.*

Que deviennent les corrélations entre ces trois instances, théorie – modèle – calcul, dans le paradigme de la donnée? La question est difficile et peut être abordée à plusieurs niveaux. Nous pensons qu'on peut distinguer deux points de vue complémentaires:

– le paradigme des grandes bases d'information
– le paradigme de la donnée, proprement dit.

4.1. *Le paradigme des grandes bases d'information.*

Selon ce premier point de vue, le paradigme de la donnée peut se comprendre comme un nouveau type d'outils de calcul qui viennent enrichir noter palette d'instruments pour mettre en œuvre les modèles. Cette vision est commandée par la massification des informations et la capacité nouvelle de les rassembler, stocker et traiter selon de mêmes traitements. Cependant, les données considérées dans ce cadre vérifient les propriétés suivantes:

– elles sont en général de même nature, c'est-à-dire qu'elles possèdent la même origine causale, et renvoient aux mêmes entités dont elles sont une trace et qu'elles permettent donc d'informer;
– le format imposé qui permet de les homogénéiser pour le calcul introduit un biais uniforme et n'altère pas leur interprétabilité au delà des difficultés habituelles de la théorisation et modélisation scientifiques.

Selon nous, il n'y pas lieu de parler de sciences des données en tant que telles pour ce cas. Cela ne signifie pas que cette approche soit sans intérêt. En effet, la capacité nouvelle de considérer un ensemble important de donnée sur un même sujet, voire l'ensemble des données, peut modifier considérablement le regard que l'on peut porter sur l'objet des données. En effet, il devient possible de dégager les propriétés de l'ensemble des objets, au lieu de n'avoir que l'ensemble des propriétés des objets. Ce changement de posture peut faire rupture. Mais il ne s'agit pas d'un changement de paradigme: le cadre est toujours celui d'une théorie permettant d'articuler les informations que l'on a sur des objets; simplement au lieu d'avoir des calculs qui ne font qu'opérationnaliser les lois théoriques, on dispose à présent d'outils de calcul qui tirent profit de la présence de données

ESL053_Azimuth_7.indd 24 06/06/16 12:15

reflétant le comportement d'objets obéissant à de mêmes lois et principes. Si cela peut changer en profondeur notre compréhension d'un domaine, cela ne change en rien notre manière de l'envisager, à savoir comme étant constitué d'objets aux comportements homogènes car réglés sur les mêmes principes.

4.2. *Le paradigme des données.*

Le paradigme des données repose sur d'autres principes que le précédent. En effet, si on reprend les principes de fonctionnement des approches fondées sur les données, comme Lev Manovich les énonce par exemple[12] 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,* cette dernière notion renvoyant au fait que les données peuvent inexactes, incertaines dans la mesure où leur masse compense par un effet statistique les erreurs ou aberrations.
– 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)[13].

Selon nous, cette présentation montre bien une caractéristique essentielle des mégadonnées, à savoir que, outre leur masse importante, elles sont de nature hétérogène et dynamique. C'est là en effet qu'on peut envisager un changement de paradigme, car il y a une rupture dans l'enchaînement classique que nous avons proposé entre la théorie, le modèle et le calcul.

Pour le ramasser en une formule, le paradigme des données ne retient que le calcul en faisant abstraction de la théorie, et donc en proposant des modèles fonctionnant comme des 'boîtes noires', c'est-à-dire proposant

[12] L. Manovich, *Cultural analytics: analysis and visualisation of large cultural data* sets, La Jolla CA, Software Studies Initiative, 2008.

[13] B. B. Bederson – B. Schneiderman, *The Craft of Information Visualization. Readings and Reflections*, Morgan Kaufmann, San Francisco, 2003.

d'agir sans intelligibilité associée. Il ne s'agirait pas là d'une perte, mais plutôt le symptôme du fait que désormais, la théorie est inutile pour réduire la complexité car le calcul, par l'efficacité des algorithmes, la puissance des machines, la masse des données, suffit pour la surmonter.

C'est donc là qu'il faut voir selon nous l'innovation, mais aussi la question, des sciences des données: le fait que la théorie soit devenue inutile, que comprendre n'est pas une étape nécessaire pour l'action. La formulation est certainement caricaturale car elle laisse à penser que cette posture est inepte. Même si nous ne la partageons pas, il nous semble qu'il ne faut pas écarter d'un revers de main cette nouvelle proposition, car elle renvoie à une tension séculaire entre le faire et le comprendre même si elle le formule en des termes bien évidemment nouveaux.

C'est un lieu commun que de constater que *l'homo faber* que nous sommes en encore, sait faire des choses qu'il ne comprend pas ou, dit autrement, il sait faire des choses sans savoir ce qu'il fait (ce que souligne par exemple Hannah Arendt dans *La condition de l'homme moderne*[14]. On peut le déplorer, mais aussi constater qu'il y a peut être plus de savoir et de sagesse dans l'action que dans la contemplation. Que la *theoria* (contemplation, selon l'étymologie du terme) ne soit pas le meilleur accès à la connaissance, et qu'il faille agir pour explorer, n'est pas une idée aussi absurde que cela. Elle mérite donc qu'on s'y arrête.

4.3. *Abstraire, faire.*

L'abstraction qui permet la construction théorique, par l'idéalisation et la généralisation, est un moyen privilégié pour dépasser la variabilité de nos sensations, la contradiction apparente de nos expériences, et construire une cohérence que l'on peut projeter en retour sur le tissu du réel. Cette abstraction est nécessaire dans la mesure où elle réduit la masse des faits tout en les conservant dans la loi qui les décrit. Comprendre a toujours été une voie essentielle pour soulager sa mémoire : comprendre vaut mieux que tout retenir.

Mais qu'en est-il de la situation où nous avons la possibilité de tout retenir via une instrumentation adaptée? La nécessité d'abstraire pour réduire la masse n'en est plus une. Le détour par la théorie devient un luxe, par conséquent inutile même si on peut la rechercher par désir esthétique ou cognitif.

De surcroît, la tradition épistémologique des siècles derniers milite également en ce sens : en voyant dans les lois scientifiques des moyens commodes

[14] H. Arendt, *Condition de l'homme moderne*, Paris, Calmann-Lévy, 1961.

ESL053_Azimuth_7.indd 26 06/06/16 12:15

pour établir des relations entre les faits d'observation ou encore les sensations, les lois n'ont d'autre légitimité que la réduction de la complexité ; par exemple, Ernst Mach, physicien positiviste inspirateur du Cercle de Vienne écrit dans *La connaissance et l'erreur*: «C'est des sensations et de leurs connexions que naissent les concepts, dont le but est de nous mener, par la voie la plus courte et la plus facile, à des idées sensibles qui s'accordent au mieux avec des sensations. L'intellection part ainsi toujours des perceptions sensibles et y retourne».

Par ailleurs, la tradition formaliste des mathématiques milite également pour une autonomie de la manipulation des symboles vis-à-vis d'une interprétation préalable de ces derniers: le sens des symboles est immanent à leur manipulation, elle est structurelle, et non substantielle. Il ne s'agit en effet pas de trouver via la formalisation théorique la meilleure expression d'objets qu'on connaît déjà ou par ailleurs: la formalisation dégage elle-même, par le jeu des relations établis par les axiomes le sens des symboles mobilisés. C'est donc la structure qui prescrit le sens des symboles qu'elle mobilise, en trouvant ensuite quels objets peuvent correspondre à ce sens, si bien que il n'est pas nécessaire ni problématique de ne pas en trouver immédiatement. Parfois, il faut attendre un certain temps pour trouver un sens mathématique ou physique aux formalisations proposées. Mais on demandera à ces symbolisations d'être précises et rigoureuses: elles doivent être consistantes (les axiomes n'entraînent pas de contradictions), complètes (ajouter un axiome introduit une contradiction), et catégoriques (les interprétations instanciant les axiomes, les modèles dans la terminologie de la logique mathématique, doivent être isomorphes entre eux). Bruno Leclercq fait ainsi remarquer:

Contrairement à Frege, Hilbert pense que la logique et les mathématiques sont moins une langue qu'une syntaxe. Il évacue dès lors la question du fondement sémantique du choix des axiomes et des règles de déduction. Un système axiomatique est d'abord un système formel, dont les symboles peuvent dans un second temps recevoir un sens en fonction des applications qu'on veut faire du système. Le choix des axiomes et des règles de déduction n'est pas dicté par quelque 'nature des choses' que ce soit, mais est au départ arbitraire, avec pour conséquence qu'une infinité de systèmes axiomatiques sont théoriquement possibles, dont cependant certains seulement pourront recevoir des interprétations utiles[15].

[15] B. Leclercq, *Intuition et déduction en mathématiques: retour au débat sur la «crise des fondements»,* Bruxelles, EME Editions, 2015, p. 209.

Cette attitude est un trait profond de notre modernité scientifique et l'épistémologie issue du Cercle de Vienne saura s'en souvenir.

Plus proche de nous, on peut également citer le «quatrième paradigme» publié par Microsoft, avec la transcription de la conférence de Jim Gray qui expose ce qui selon lui retrace l'histoire des sciences jusqu'aux sciences des données[16]:

– la science empirique: on décrit les phénomènes naturels et on reproduit des procédés techniques qui n'ont pas forcément d'explication théorique, tant qualitative que quantitative;
– la science théorique: les lois formalisent les phénomènes naturels qui en deviennent les instanciations;
– la science computationnelle: le calcul scientifique permet d'opérationnaliser les lois et les modèles et par conséquent de simuler des phénomènes complexes;
– la science des données: elle unifie la théorie, l'expérience et la simulation; elle récupère des données engendrées par des capteurs ou simulées par des modèles, données qui sont ensuite traitées en transformées en informations ou connaissances; ces dernières sont exploitées par des analystes s'appuyant sur les statistiques, le management et la visualisation des informations.

Dans cette vision, les sciences des données sont autant une coupure vis-à-vis de la théorie et ses modèles qu'une récapitulation qui les assume sans avoir à les expliciter en tant que tels. Cependant, il est clair que la collecte des données rend inutile l'abstraction de la formalisation puisque le calcul permet de les manipuler et de les transformer. Cette approche confirme également ce que nous suggérions ci-dessus: le traitement calculatoire s'effectue de manière systématique et homogène entre toutes ces données, indépendamment de leur nature, histoire et signification; en particulier il n'est pas nécessaire de s'assurer que le traitement effectué est commensurable avec la manière dont la donnée est obtenue: il est inutile que l'appareil calculatoire soit homogène avec l'appareil de mesure, c'est-à-dire que ces appareils renvoient aux mêmes soubassements théoriques. On peut donc conclure que le paradigme de la donnée est une rupture dans la mesure où elle rompt avec la formalisation théorique qui ajuste dans une même intelligibilité l'obtention

[16] T. Hey – S. Tansley – K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Redmond, Microsoft Research, 2009.

ESL053_Azimuth_7.indd 28 06/06/16 12:15

de la donnée et son traitement; le traitement de la donnée est désormais le seul lieu d'élaboration de la théorie et de l'expérience.

4.4. *Faire, comprendre, faire comprendre.*

La principale question qui se pose alors est savoir quel statut donner aux résultats obtenus par une telle manipulation de données qui ne s'appuie pas sur une conception théorique sous-jacente. On a souvent critiqué la notion de 'données' en arguant qu'il faudrait davantage les appeler des 'obtenus' ou des 'construits': cela est évidemment exact dans la mesure où ces données ne viennent pas de nulle part. Cependant, ce n'est pas de cela qu'il s'agit ici: la donnée est une donnée par on ne s'occupe pas de la manière dont elle est obtenue ou construite: certes, elle est obtenue et construite selon certaines procédures; mais on l'ignore délibérément car ce n'est pas pertinent. Par conséquent, cette critique, aussi juste soit-elle, manque-t-elle son objet car elle se méprend sur ce dont il est question: comme nous l'avons montré plus haut, la science des données sait très bien que les données sont issus de robots, de serveurs, d'appareil de mesure, de simulateurs, etc. Mais elle sait ici qu'elle peut l'ignorer car les traitements qu'elle effectue sont indépendants de ces procédures. C'est donc la grande force de ce paradigme et non sa faiblesse.

Cependant, il nous semble que ce paradigme bute sur une difficulté majeure qui est celle de l'intelligibilité. Par intelligibilité, nous comprenons le fait de rendre compréhensible ce qu'on considère par la production d'un concept, d'une raison. L'intelligibilité, c'est le fait de rendre raison de quelque chose, d'en donner la raison.

Pourquoi après tout faut-il comprendre et donner la raison des choses ou de ce qui arrive? Quelle différence y a-t-il entre le pouvoir de prédire ce qui arrive même si on ne comprend pas le sens de ce qui arrive, et le fait de comprendre ce qui arrive même si on ne sait pas dire exactement comment?

Selon nous, la compréhension revient à pouvoir établir un lien entre les relations entre des événements d'une part et les langages dont nous disposons (langues naturelles, langues formalisées) d'autre part. Autrement dit, le langage est le lieu de la compréhension car il permet de reprendre dans la discursivité, sous la forme de liens logiques, le constat empirique de ce qui arrive et le lien entre les phénomènes.

Le problème que pose les sciences des données est qu'elles rompent le lien entre discours et faits, entre langage et événement, entre logique et phénomène. En rompant le lien traditionnel entre la théorie et la mesure, faisant de cette dernière une donnée indépendante et autonome vis-à-vis de

son origine, les sciences des données proposent des traitements qui, même s'ils renvoient à des outils mathématiques perfectionnés, sont alors sans lien avec ce dont parlent les données et ce dont elles sont les données.

Cette perte d'intelligibilité se retrouve dans la difficulté constatée à interpréter les résultats produits: la visualisation des informations, la présentation des résultats renvoient à des interprétations arbitraires, arbitraires car elles ne sont plus conditionnées par le lien organique entre le réel sous-jacent aux données, les méthodes d'obtention de ces données, leur traitement et donc le résultat de ces traitements. Puisqu'il n'est plus possible de se raccrocher au statuts propre des données (qui ne sont plus des mesures résultants d'appareils eux mêmes issus d'appareils théoriques), l'expert analysant les données devient l'interprétant unique (au sens peircien) des résultats: il doit donc retrouver dans ce qu'il considère, des motifs et des signifiances qui, puisqu'elles ne peuvent venir de l'ordre théorique aux données, viennent de ce qu'il sait déjà. On en arrive au paradoxe que les sciences des données ne peuvent nous apprendre que ce que l'on sait déjà, car c'est la seule intelligibilité qu'on peut leur conférer.

C'est donc le paradoxe de Ménon de Platon qui se rejoue ici: on sait que dans ce dialogue, Platon montre que la seule explication possible à notre capacité à apprendre des choses nouvelles est que nous les sachions déjà. Car il est impossible d'apprendre quelque chose de nouveau; en effet, dans une situation donnée,

– si on se retrouve devant une connaissance nouvelle, on ne peut la reconnaître comme telle, car elle est nouvelle et par conséquent on ne la connaît pas déjà;
– si on trouve une connaissance, c'est qu'on la connaissait déjà puisque on a pu la reconnaître.

Mutadis mutandis, l'analyse des résultats produits par la science des données reproduit ce paradoxe: puisque la donnée est utilisée comme un absolu et non comme un obtenu, on construit un ordre nouveau incommensurable grâce aux traitements effectués, si bien que ces traitements nous apprennent quelque chose que si on sait reconnaître dans cet ordre nouveau ce que nous savions déjà sinon on ne sait pas donner un statut ni une raison à ce que l'on constate.

Autrement dit, puisque l'intelligibilité ne peut venir du réel car la donnée s'en est coupée, elle ne peut venir que de l'interprétation, qui puise dans ses propres interprétants et connaissances préalables pour donner du sens à ce qui est trouvé.

ESL053_Azimuth_7.indd 30 06/06/16 12:15

Il y a donc bien un prix à payer dans la rupture qu'opèrent les sciences des données. En permettant de rapprocher des données hétérogènes dans des traitements homogènes, ces sciences ouvrent des perspectives calculatoires totalement nouvelles, mais elles renvoient à l'arbitraire de l'interprétation de leur résultat. L'intelligibilité perdue de la donnée, qu'on ne peut plus interroger depuis son origine, implique un arbitraire de l'interprétation des résultats produits: soit on débouche sur une démarche esthétique où la production de sens est constatée mais dans une irruption constatée et sans lien plausible avec les données d'origine, soit on reconnaît des motifs signifiants renvoyant à des connaissances préalables.

5. *De nouveaux enjeux.*

Si l'intelligibilité perdue est le problème rencontré par les sciences des données, cela ne permet pas de les condamner, mais simplement d'indiquer là où il faut travailler pour les acclimater et apprendre à nous en servir. Comme souvent, le problème essentiel est de parvenir à surmonter la complexité des faits, complexité provenant de leur masse et de leurs interdépendances. La voie traditionnelle fut l'abstraction et la théorisation, la voie contemporaine est le calcul et la manipulation des données. La voie traditionnelle ne dit pas quoi faire qu'on ne sait pas théoriser, la voie contemporaine ne dit pas comment interpréter ou comprendre les résultats qu'on a obtenus par le traitement. Quand la complexité déborde nos capacités formalisatrices, la voie traditionnelle ne nous permet que de constater notre impuissance à faire quelque chose des données disponibles. La voie contemporaine suggère une manière de faire en l'absence de théorie, mais bute sur une incapacité à comprendre.

Exposé ainsi, on comprend qu'il n'est pas question de s'abstenir de la nouvelle puissance de faire que nous procurent ces nouvelles approches, d'autant que la complexité instaurée par la numérisation de monde (naturel et culturel) ne peut que se renforcer. Il faut donc travailler à conjuguer la puissance du faire et la puissance de comprendre.

On ne peut qu'émettre ici des orientations et recommandations. Celles-ci doivent s'appuyer sur notre diagnostic concernant l'origine des difficultés constatées:

– l'oubli de l'origine de la donnée;
– la formatage uniformisant des données pour les rendre manipulables via de mêmes outils;
– des traitements qui n'ont a priori aucun lien avec les méthodes et outils ayant permis d'obtenir les données.

– une présentation des résultats qui les montrent dans une organisation visuelle et sémiotique indépendantes de la nature des données et de leur signification;
– une analyse des résultats qui renvoie à un arbitraire interprétatif dans la mesure où les seuls interprétants sur lesquels l'analyste peut s'appuyer sont ses propres connaissances.

On peut donc dès lors proposer les suggestions suivantes:
– l'articulation entre le format et la nature de la donnée:
  › le format est souvent suggéré par le type de traitement que l'on veut effectuer: des caractères pour construire des grammaires, des matrices de nombres pour des traitements statistiques, etc.; la question du format est souvent considérée comme secondaire car notre héritage de l'épistémologie de la mesure nous a habitué à manipuler des données formalisées via leur mode d'obtention dans une forme adéquate aux traitements subséquents. Or, dès lors que les données sont hétérogènes, l'articulation du format les unifiant pour leur appliquer des traitements calculatoires et de leur nature devient une question essentielle et première.
  › L'enjeu est alors une épistémologie de la donnée, succédant à une épistémologie de la mesure, où un même cadre théorique d'intelligibilité permet de rapporter le format de la donnée calculée au format de la donnée obtenue.
  › Cet enjeu n'est pas nouveau: tout chercheur en sciences sociales effectue ce travail préalable à sa recherche; mais les sciences des données s'appuient sur des approches automatisées de captations de données et de rassemblement en de grandes bases; l'étape méthodologique de recollement entre formats est alors ignorée; l'enjeu est donc d'introduire dans les procédures des critères de compatibilité entre données collectées et données calculées.
– Traitement des données et présentations des résultats:
  › Les sciences des données se sont popularisées par des présentations de résultats suggestives et saisissantes, donnant à voir (plus qu'à comprendre d'ailleurs) des phénomènes qu'on n'appréhendait que difficilement jusqu'alors: cartes de Web, relations entre des phénomènes apparemment indépendants, etc. On apprend à voir grâce à elles des aspects insoupçonnés du réel. Macroscope de la globalité pour reprendre le terme suggestif de Joël de Rosnay[17], ces visualisations dévoilent de nouvelles réalités.

[17] J. De Rosnay, *Le macroscope: vers une vision globale*, Paris, Seuil, 2014.

› La principale difficulté est que les visualisations sont davantage conçues pour donner à voir que pour refléter les résultats: optimisées pour la perception, elles suggèrent plus qu'elles ne démontrent. En particulier, un analyste est incapable de faire le lien entre ce qu'il constate et le traitement sous-jacent des données. Cette opacité interprétative est alors compensée, comme on l'a dit, par la reconnaissance de savoir déjà-là.
› La suggestion serait alors de permettre de naviguer des traits signifiants de la présentation des résultats aux traitements sous-jacents, jusqu'aux données associées. Puisque les traitements sont complexes, et les données nombreuses, cela exige d'inventer des modes de présentation intermédiaire car il sera la plupart du temps difficile et impraticable de revenir à la donnée singulière; ces visualisations intermédiaires ont pour vocation de montrer le rôle et la contribution de groupes de données, et ainsi d'avoir une présentation progressivement analytique des visualisations globales.

– Corrélations et explications
› Finalement, l'intelligibilité repose sur la capacité entre rendre raison des corrélations constatées. Rendre raison signifie construire un ordre logique (via un langage) reflétant un ordre phénoménal ou, ici, un ordre corrélatif entre données.
› Au delà de l'articulation entre le format de la donnée obtenue et celui de la donnée calculée que nous avons évoquée plus haut, la question est d'expliciter le savoir que nous avons déjà sur les données que nous mobilisons. Là encore, nous sommes piégés par notre traditionnelle épistémologie de la mesure où cette dernière est obtenue dans un contexte où elle est d'emblée positionnée et intégrée dans un édifice théorique. L'épistémologie de la donnée porte sur des informations aux origines diverses, parfois douteuses, si bien qu'elles ne portent en elles aucune valeur théorique particulière (des factures de magasins par exemple, utilisées par les sciences des données pour le marketing).
› Cette faiblesse théorique de la donnée, que les sciences des données dépassent en leur imposant un traitement massif et homogène ignorant tant leur origine que leur incertitude, ne peut être oubliée: il faut réintroduire le fait d'expliciter des hypothèses, fussent-elles de sens commun quand on aborde des données issues du quotidien (grande consommation) ou de domaines spécialisés mais faiblement théorisés (économie). Ces hypothèses fonctionnent comme des heuristiques: elles guident l'investigation d'une part, et fournissent un cadre interprétatif d'autre part.

L'enjeu des sciences des données est par conséquent de réintroduire une intelligibilité sans avoir à perdre ce qu'elles nous ont permis de gagner, à savoir une puissance opératoire intégrant des données massives et hétérogènes, que nous ne savions pas gérer auparavant. Au delà d'un nominalisme réducteur où les faits culturels sont rapportés à des données amnésiques de leur nature et origine pour dégager leur sens uniquement à partir du jeu de leurs relations imposées par le formatage effectué, il convient de rapporter les traitements à d'une part une intelligence du rôle des données dans la production du résultat d'une part et d'autre part à une conception théorique articulant les données au réel qui les produit, renvoyant, lâchons le mot, les corrélations à des causes.

L'enjeu est par conséquent de permettre de rapporter la perception de la globalité à la contribution du local, de confronter le sens perçu grâce à la présentation des résultats au rôle des données dans cette construction. Discutant les bases documentaires, Moretti[18] introduit les notions de «lecture distante» et de «lecture proche»: il ne s'agit pas en effet de substituer à l'approche traditionnelle de lecture d'un document pour lui-même une lecture globalisante s'appuyant sur le traitement de la base documentaire dans son ensemble; l'enjeu est de composer ces deux types de lecture en permettant des aller et retours entre *distant reading* et *close reading.* Mais dans notre cas, il conviendrait davantage de parler de «perception globale» d'une part et de «travail analytique» d'autre part: en effet, la production de sens reposant principalement sur des visualisations de résultats, on a souvent une perception qualitative d'une globalité; par ailleurs, la donnée ne prête pas à une «lecture», mais bien plutôt à un travail où l'analyse (le fait de décomposer et d'opposer les fragments obtenus) permet de comprendre l'apport, le rôle et la signification de la donnée.

Pour que ce travail soit productif, il est nécessaire que l'analyse soit conduite en faisant appel aux heuristiques et hypothèses formulées sur la nature des données, leur rôle causal, les relations entre elles. Les sciences des données, apportant une nouvelle manière de produire du sens, appellent naturellement comme l'une de leur dimension, ce travail analytique, fondé sur la causalité et l'explication, rendu nécessaire par l'exigence de rendre raison du sens constaté.

Conclusion

Les sciences des données constituent une rupture épistémologique. Elles consacrent le fait que nous passons d'une ère de la mesure à celle de la

[18] F. Moretti, *Graphes, cartes et arbres: modèles abstraits pour une autre histoire de la littérature*, Paris, Les Prairies ordinaires, 2008.

ESL053_Azimuth_7.indd 34 06/06/16 12:15

donnée, reconfigurant ainsi les rôles respectifs de la théorie, du modèle et du calcul. Fondées sur la possibilité créée par le numérique de rassembler des données hétérogènes en une synthèse calculante qui les homogénéise, ces sciences proposent des synopsis calculées dont il faut rendre raison. L'intelligibilité devient la clef qui permettra aux sciences des données de devenir une réalité et pas seulement une promesse.

Alors que pour le moment, l'exploitation des résultats reconduit le paradoxe du Ménon où on ne peut découvrir que ce que l'on savait déjà, la complexité des représentations et leur apparence (apparent) arbitraire vis-à-vis des données initiales impliquant que le seul interprétant possible étant les connaissances préalables de l'analyste, l'enjeu est de permettre un travail analytique où la donnée reprend son rôle d'interprétant via les hypothèses et heuristiques qu'on formule à son endroit quant à sa nature, son rôle causal, les relations avec son contexte d'obtention et les autres données.

L'intelligibilité, où le langage permet de rendre raison des phénomènes et de leur relations, reste donc la clef. Mais faut-il s'en étonner?

Bruno Bachimont
Sorbonne Universités – Université De Technologie de Compiègne