Les *big data* dans les humanités

Viktor Mayer-Schönberger et Kenneth Cukier *Big Data* *A Revolution That Will Transform How We Live, Work, and Think*	Londres, John Murray, 2013, 256 p.
Jaron Lanier *Who Owns the Future ?*	Londres, Penguin, 2013, 400 p.
Erez Aiden et Jean-Baptiste Michel *Uncharted* *Big Data as a Lens on Human Culture*	New York, Riverhead Books, 2013, 288 p.

Lorsqu'on soumet au moteur de recherche Google l'expression *big data*, plusieurs centaines de millions de pages sont proposées comme résultat. Un tel succès interdit que l'on se moque ! S'il n'avait pas accédé à une si grande renommée, le terme *big data*, littéralement « grosses données » ou, plus communément, masses de données, prêterait à sourire. Pour saisir ce qui fait la nouveauté et l'importance du concept et en évaluer les conséquences économiques, épistémologiques, éthiques et politiques, nous partirons de trois ouvrages récents [1]. Nous insisterons surtout

1. *Big Data. A Revolution* et *Who Owns the Future ?* sont disponibles en traduction française : *Big Data. La révolution des données est en marche*, Paris, Robert Laffont, 2014 ; *Internet : qui possède le futur ?*, Paris, Le Pommier, 2014. À ceux que n'effarouchent pas

sur les conséquences d'ordre épistémologique du traitement des « grosses données » dans les humanités, mais aussi dans les sciences de la nature. En effet, c'est au regard de l'évolution de ces sciences vers les *e-sciences* qu'on peut le mieux appréhender l'évolution des humanités vers les humanités numériques. Encore faut-il d'abord caractériser ces « grosses données » : d'abord au plan quantitatif, puisque leur ampleur semble sans commune mesure avec celle des données classiques, puis au plan qualitatif puisque le traitement de masses de données est réputé différer de la simple exploitation de contenus numérisés. Enfin, nous conclurons sur une évocation des conséquences économiques, éthiques et politiques des *big data*.

Que signifie big data *?*

On l'a dit, la notion de données n'est pas neuve et leur recueil systématique est lui-même très ancien. Comme le rappelle Viktor Mayer-Schönberger, dès la plus haute antiquité, au temps des Assyriens, on procéda à des statistiques et la Bible évoque le recensement de la population du royaume qu'ordonna David en vue de lever l'impôt. Depuis l'avènement des premiers ordinateurs, les données représentées sous forme numérique binaire, comme des suites de 0 et de 1, se manipulent de façon automatique. Cette numérisation ne se limite pas à des données quantitatives : textes, images, sons, vidéos, voire sensations kinesthésiques se traduisent en flux d'information et se traitent comme tels. Là encore, rien de vraiment nouveau : il y a bientôt soixante-dix ans, Vannevar Bush, Alan Turing, Norbert Wiener et d'autres en eurent l'intuition. Bref, les notions de données, de recueil systématique, de numérisation et de traitement automatique apparaissent comme anciennes.

L'innovation tient à l'automatisation et à la massification de la collecte par le truchement de machines. Aujourd'hui,

---

les équations, nous conseillons aussi la lecture d'un numéro spécial de la revue du laboratoire Lincoln du MIT intégralement consacré aux aspects techniques des très grands graphes : *Lincoln Laboratory Journal*, « Graphs and Networks », vol. 20, n° 1, 2013, www.ll.mit.edu/publications/journal/journal.html#currentjournal.

des capteurs robotisés enregistrent et stockent des données en quantité. Des observations de toutes sortes – images, sons, températures – se font sans l'intervention de l'homme. À cette automatisation de la saisie s'ajoute la production d'annotations collectives centralisées grâce à l'interconnexion de la population sur le Web. Ainsi, ce qu'on appelle le *crowdsourcing* – littéralement la « collecte par la foule » –, met à contribution de très nombreuses personnes qui collaborent, consciemment ou inconsciemment, à ce recueil de données. À titre d'illustration, le logiciel Waze enregistre, chaque fois qu'il est connecté, la vitesse et la position des véhicules sur lesquels il se trouve embarqué, et les transmet à une centrale qui les agrège pour en déduire la densité de circulation. De même, il existe des sites de science collaborative[2] qui engagent tous ceux qui le souhaitent dans des activités scientifiques d'interprétation, par exemple, pour caractériser la forme des objets célestes. D'autres sites proposent d'annoter des tableaux[3], voire des textes. Chaque requête sur un moteur de recherche comme Google (et *a fortiori* chaque achat) est automatiquement exploitée.

Prenons l'exemple de la numérisation des textes. Jusqu'à présent, celle-ci se faisait manuellement, avec éventuellement l'aide de logiciels de reconnaissance optique de caractères (*Optical Character Recognition*), mais à l'échelle limitée d'une bibliothèque. Depuis une dizaine d'années, avec le projet *Google Books*, ce sont des réseaux d'acteurs, dans l'ensemble des bibliothèques du monde, qui sont sollicités pour contribuer à ce grand projet de numérisation. Il résulte de cette massification de l'acquisition des données, produite grâce aux machines et à leur mise en réseaux, un changement de dimension que traduit la formule « grosses données ».

Ordres de grandeur

Pour nous faire une idée du changement quantitatif, donnons quelques ordres de grandeur en prenant pour référence deux unités bien connues : le livre et la bibliothèque. Si l'on considère, en faisant abstraction des images, qu'un

2. https://www.zooniverse.org/.
3. http://tagger.thepcf.org.uk/.

livre compte à peu près un million de caractères typographiques – estimation plutôt généreuse –, il « pèse » au moins un million d'octets d'information, c'est-à-dire un mégaoctet (Mo = 10^6 octets). Le catalogue des livres et imprimés de la Bibliothèque nationale de France contient à peu près quatorze millions d'ouvrages. En reprenant cette estimation de 10^6 octets pour un livre, la BNF correspond donc à peu près à 14×10^{12} octets, soit quatorze téraoctets (To = 10^{12} octets). Ce chiffre assez conséquent correspondait pour l'homme éclairé du XXe siècle à l'horizon ultime du savoir. Mettons-le en perspective avec les capacités de stockage actuelles : un disque dur externe de quatre téraoctets vaut environ cent cinquante euros aujourd'hui. Il est donc désormais possible de posséder un équivalent numérique de la BNF chez soi, à moindre coût (moins de cinq cents euros), somme dérisoire comparée au prix de ce grand bâtiment ! Quant au volume total du Web, il a été évalué en 2012 à environ 2,8 zettaoctets (Zo = 10^{21} octets), ce qui fait 2,8 milliards de téraoctets, c'est-à-dire à peu près deux cents millions de BNF.

Considérons à présent le rythme de la production de nouvelles données, en gardant la BNF pour unité de compte. Aujourd'hui, les messages publiés sur Twitter, pourtant limités chacun à cent quarante caractères, représentent 7 To par jour, soit une demi-BNF quotidienne. Le radiotélescope Murchison Widefield Array en Australie produit 7 000 To par minute de données brutes, c'est-à-dire cinq cents BNF. Après traitement et filtrage, il reste 50 To par jour, c'est-à-dire un peu plus de trois BNF. Selon les prospectivistes contemporains, le Web représentera environ 7,9 Zo en 2015, soit un peu plus d'un demi-milliard de BNF. Et cette croissance-là n'est pas prête de s'arrêter.

Ces quelques chiffres aident à se faire une idée grossière des caractéristiques quantitatives de ce que l'on appelle les *big data*. Mais le fait d'emmagasiner de grandes masses d'information n'est pas leur seule caractéristique. On les définit souvent par la formule des « 3 V » – pour *volume*, *vélocité* et *variété*. Le volume, c'est-à-dire la quantité proprement dite, oscille entre le téraoctet et le pétaoctet (Po = 10^{15} octets = 1000 To), à savoir, entre un dixième de BNF et cent BNF. La vélocité renvoie au fait que cette masse de données se renouvelle en permanence. Enfin, les données sont variées au sens où elles

sont hétérogènes : elles peuvent contenir du texte, des images, des sons, etc. Le texte lui-même peut intégrer différentes langues, divers systèmes d'abréviations, etc.

Mesurées à l'aune des « grosses données », les études littéraires portent sur des volumes bien faibles. L'intégralité des corpus des auteurs classiques, même en considérant les brouillons, les œuvres critiques, les articles de journaux et les correspondances, se trouve incluse dans une petite partie de la BNF. Quant à la vélocité et à la variabilité, elles apparaissent quasi nulles dans le cas des données littéraires. Les archives historiques ne semblent pas beaucoup plus grosses, ni plus « véloces », même si elles sont un peu plus variées. Dans les sciences de l'homme et de la société, il n'y a guère que la psychologie et la sociologie qui rassemblent de gros corpus : la première, parce qu'elle enregistre l'ensemble des données physiologiques des individus prises avec toutes sortes d'appareils de mesure, par exemple des oculomètres qui enregistrent les mouvements des yeux ; la seconde, lorsqu'elle explore le graphe du Web, les réseaux sociaux, les outils de microblogage, comme Twitter, et les échanges téléphoniques. Portant sur des volumes plus petits, mais assez conséquents au regard de ceux que l'on exploitait précédemment, les études historiques et littéraires bénéficient pourtant elles aussi de l'apport des techniques de traitement de l'information. L'ouvrage de Erez Aiden et Jean-Baptiste Michel montre très en détail ce que la variation temporelle de la fréquence des mots isolés, ou des séquences de n mots consécutifs (ce que l'on appelle en termes techniques des *n-grammes*), peut révéler de l'évolution des mentalités. On peut en donner un exemple très simple : si l'on compare sur un graphique l'évolution temporelle de la fréquence d'emploi des deux séquences de quatre mots (*4-grammes*), « The United States is » et « The United States are » dans la littérature américaine, on voit progressivement naître et s'imposer la première formule aux dépens de la seconde, ce qui atteste de l'émergence progressive d'une unité nationale aux États-Unis durant le XIXe et le début du XXe siècle.

Du quantitatif au qualitatif

Au-delà de ces caractéristiques chiffrées, on a pu dégager des spécificités qualitatives qui dérivent en grande partie

de la collecte à la fois massive et automatisée (ou du moins machinique) des données.

Comme la collecte se fait de façon automatique, elle n'obéit pas à un protocole précis conçu pour valider ou invalider une théorie. Il en résulte un certain désordre dans ces données engendrées de façon aléatoire. Pensons à une étude médicale destinée à évaluer les effets de tel ou tel médicament dans le traitement de telle maladie. Les approches classiques procèdent en double aveugle avec deux substances, l'une active comprenant le médicament que l'on souhaite évaluer, l'autre correspondant soit à un placebo, soit à un autre médicament dont on connaît les effets positifs. On considère alors un ensemble représentatif de malades et on tire aléatoirement ceux qui reçoivent la substance active que l'on cherche à tester et ceux qui reçoivent soit le placebo, soit la substance à comparer. Avec les « grosses données », on ne prend pas ces précautions méthodologiques : tous les patients atteints de la maladie sont systématiquement enregistrés. On filtre ensuite ceux qui ont pris le médicament en question et les autres, puis on effectue des tests statistiques. Notons que les cas étant recueillis à la volée, il se peut que certains d'entre eux soient « bruités », en l'occurrence, que certains patients n'aient pas pris leur traitement de façon régulière, ou qu'ils aient arrêté de se soigner, ou encore que l'on ait mal enregistré les données. Une étude clinique classique les exclurait du protocole. Avec les « grosses données », on minimise l'impact de ces imperfections par le nombre : celui-ci étant considérable, on parvient toujours à extraire des corrélations significatives.

Ajoutons que, dans les études statistiques classiques, on dispose d'un modèle des données pour fabriquer un échantillon représentatif de la population à étudier. Si l'on fait une étude clinique, on sera capable avec ce modèle d'établir la répartition des patients à inclure dans le protocole, de façon à empêcher les biais et à obtenir un résultat probant. De même, lorsque l'on fait un sondage, le modèle conduit à inclure une certaine proportion de jeunes, de personnes âgées, d'ouvriers, de cadres, etc., dans l'échantillon. Avec les « grosses données », on ne dispose d'aucun modèle *a priori* de la population puisque l'on considère l'ensemble des données. On ne dispose pas non plus de théorie initiale que l'on souhaite valider ou invalider, puisque la collation se fait

systématiquement. Illustrons cela par un exemple qui revient dans tous les ouvrages ici traités. La société Google établit systématiquement des corrélations à partir des requêtes soumises à son moteur de recherche. C'est ainsi qu'elle a détecté une épidémie de grippe naissante dans une ville des États-Unis, parce qu'une partie anormalement élevée de la population cherchait des informations sur le traitement de symptômes s'apparentant à ceux de la grippe. Google, avant de la découvrir, n'avait aucune intention de prévoir cette épidémie de grippe. On procède donc ici par détection d'anomalies, sans savoir ce que l'on va trouver : « on part à la pêche. » Autrement dit, les « grosses données » sont à la fois exploratrices et heuristiques, là où les statistiques classiques se présentaient comme probatoires.

Au reste, soulignons que l'absence d'échantillonnage de la population ne tient pas uniquement à l'absence de modèle initial, mais aussi au principe même d'exploitation des « grosses données », qui vise à trouver des corrélations et des anomalies, tandis que les statistiques classiques cherchaient principalement à valider des causalités en mettant au jour des régularités. Si l'on avait procédé à un échantillonnage à partir des requêtes de Google, la ville où l'épidémie s'est déclenchée aurait pris une part si faible dans l'échantillon que l'épidémie aurait été indécelable. Il résulte de cette absence d'échantillonnage l'impossibilité de réduire le nombre de données par tirage aléatoire : toutes doivent être traitées. Cela veut dire que l'on a besoin d'algorithmes efficaces sur des machines puissantes.

En résumé, les « grosses données » se caractérisent par l'absence de protocole de collecte, de théorie initiale, de modèle et d'échantillonnage, par leur désordre aussi et par la capacité des méthodes qui les accompagnent à neutraliser les imperfections d'une collecte de masse.

Révolution épistémologique

Les sciences expérimentales naissent à l'âge moderne. Elles se distinguent des sciences d'observation qui avaient prévalu depuis l'Antiquité en ce qu'elles ne se contentent pas de contempler passivement la nature pour la comprendre, mais interagissent avec elle. Au lieu de partir des

phénomènes pour induire des causes probables, comme auparavant, elles partent d'une théorie qu'elles confrontent à la réalité. Cette confrontation passe par la construction de dispositifs matériels d'observation ; selon qu'il y a écart ou non entre ces observations et ce que la théorie anticipe, celle-ci se trouve invalidée ou validée. Les « grosses données » révoquent ce schéma classique de la science expérimentale : la théorie ne prime plus ; les dispositifs matériels de recueil d'observations ne sont pas conçus au regard d'une hypothèse que l'on cherche à prouver. La collation des données se fait automatiquement, sans idée *a priori*. On exploite ensuite ces données à l'aide d'outils d'intelligence artificielle et d'apprentissage machine, en procédant à ce que l'on appelle des expériences *in silico*, c'est-à-dire à des expériences conduites sur les puces de silicium au moyen de systèmes de traitement de l'information, par opposition aux expériences *in vivo*, qui se font sur le vivant, et *in vitro*, dans des éprouvettes de verre. Ces expériences testent systématiquement de grands nombres d'hypothèses sur les données. Nous assistons donc, dans les sciences de la nature, à une révolution d'ordre épistémologique qui change le statut de l'expérience. Notons toutefois que, contrairement à ce qu'on lit parfois – notamment dans *Big Data. A Revolution* et, de manière plus nuancée, dans *Uncharted* –, tout *a priori* n'est pas absent du processus d'exploitation des « grosses données ». En effet, les modalités de numérisation et les standards des données influent grandement sur les résultats produits. Les outils de fouille (c'est-à-dire d'analyse) des données explorent de vastes champs d'hypothèses qui n'émergent pas spontanément car ils dépendent, eux aussi, des catégories que l'on propose aux logiciels d'apprentissage des machines.

Si nous essayons de transposer ce schéma aux humanités, nous nous heurtons à une difficulté de taille : ces dernières ne visent ni à prédire l'inconnu, ni à construire des lois générales qui résument les observations, mais à expliquer les œuvres humaines. Elles ne procèdent pas à une généralisation des cas particuliers en effaçant leurs singularités. Elles cherchent ce qui, dans le particulier, relève d'un principe général, afin de mieux le comprendre. La logique inductive des sciences de la nature ne saurait donc

s'appliquer telle quelle. D'ailleurs, il n'existe pas d'expérimentation à proprement parler dans des disciplines comme la littérature ou l'histoire. La notion d'échantillonnage, qui passe par le tirage aléatoire d'individus représentatifs, perd son sens puisque chaque œuvre et chaque événement historique sont singuliers et, à ce titre, demeurent des objets potentiels d'étude.

Il existe cependant des approches empiriques des humanités, fondées sur l'observation et non sur une théorie générale. Rien ne s'oppose donc à ce que l'on applique les méthodes empiriques des *big data* aux disciplines des humanités, en partant des œuvres numérisées, et non d'échantillons représentatifs, pour détecter des corrélations ou des anomalies. On peut notamment proposer de nouvelles approches de l'intertextualité par une exploration systématique des citations ou des reprises entre auteurs[4]. Certaines reprises étant non délibérées, leur étude gagne à ne pas se faire sur un échantillon préétabli d'œuvres et d'auteurs. De même peut-on étudier les évolutions stylistiques au cours du temps ou repérer les structures annulaires[5] dans la littérature antique. Toutes ces études empiriques conduites avec l'aide de l'ordinateur sur la quasi-intégralité des œuvres humaines numérisées relèvent des « grosses données » en ce qu'elles ne procèdent pas à un échantillonnage, qu'elles n'impliquent ni protocole de recueil, ni théorie initiale et ni modèle de données. En revanche, ce qui semble la caractéristique la plus évidente des « grosses données », à savoir la « grosseur », fait ici défaut. Ce sont des « grosses données » un peu maigres auxquelles on a affaire dans les humanités – maigreur toute relative, cependant, puisqu'elle se mesure en centaines de milliards de caractères typographiques…

*

4. J.-G. Ganascia, P. Glaudes et A. Del Lungo, « Automatic Detection of Reuses and Citations in Literary Texts », *Literary and Linguistic Computing*, vol. 29, n° 3, 2014.

5. La notion de structure annulaire est propre aux littératures antiques et orientales : il s'agit d'une formule répétée de façon cyclique avec éventuellement de petites variations.

Les données occupent une place de plus en plus importante tant dans le champ intellectuel, où elles sont sources de connaissances, que dans celui de la finance. Elles sont réputées conditionner la richesse dans de nombreux secteurs d'activités, bien au-delà d'Internet ou des médias, comme l'attestent les capitalisations boursières impressionnantes de sociétés clefs dans la collecte des données – Google, Facebook et Amazon. Une question s'impose : qui possède ces données et qui en tire avantage ? Soulignons que l'acquisition est toujours collective : elle se fait soit à l'aide d'infrastructures coûteuses comme les grands télescopes ; soit avec le concours de réseaux d'acteurs publics comme pour le projet *Google Books* qui mobilisa de nombreuses bibliothèques ; soit en recourant à des foules d'individus volontaires ou non. Il semblerait légitime d'attribuer la propriété et le bénéfice de l'exploitation des données à ceux qui les ont produites. Certains vont plus loin et préconisent leur mise en commun. On le sait, ceux qui en profitent sont les grands acteurs qui les centralisent et les traitent. D'où de nombreuses questions d'ordre éthique. Les auteurs de *Big Data* et de *Uncharted* insistent sur les risques que pose, pour la vie privée, le fait que l'intégralité des données individuelles se retrouve entre les mains d'acteurs puissants qui pourraient être tentés d'abuser de leur pouvoir. Jaron Lanier soutient, pour sa part, que la protection de la vie privée n'est pas une fin en soi car l'individu ne s'épanouit pas dans l'isolement. Il estime même qu'à mettre trop l'accent sur cet aspect des choses, on brouille les vrais enjeux qui, selon lui, portent sur le « partage », sur ce qu'il a d'illusoire et sur son iniquité. Le vrai danger à ses yeux, d'ordre politique, est la prise de contrôle de toutes les données par une minorité de sociétés : celles que Jaron Lanier nomme des « *Siren Servers* », car ces servers, loin de nous servir nous asservissent, à l'instar des sirènes, en chantant les louanges de la gratuité, du partage et de l'échange.

Jean-Gabriel GANASCIA.