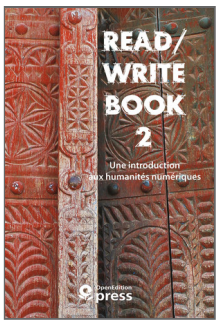

Pierre Mounier (dir.)

Read/Write Book 2
Une introduction aux humanités numériques

OpenEdition Press

Six provocations à propos des *big data*

Danah Boyd et Kate Crawford

Laurence Allard, Pierre Grosdemouge et Fred Pailler

Éditeur : OpenEdition Press
Lieu d'édition : Marseille
Année d'édition : 2012
Date de mise en ligne : 21 septembre 2012
Collection : Read/Write Book
ISBN électronique : 9782821813250

http://books.openedition.org

Édition imprimée
Date de publication : 21 septembre 2012

Référence électronique
BOYD, Danah ; CRAWFORD, Kate. *Six provocations à propos des big data* In : *Read/Write Book 2 : Une introduction aux humanités numériques* [en ligne]. Marseille : OpenEdition Press, 2012 (généré le 27 décembre 2018). Disponible sur Internet : <http://books.openedition.org/oep/273>. ISBN : 9782821813250. DOI : 10.4000/books.oep.273.

Six provocations à propos des *big data**

Danah Boyd
Kate Crawford

Traduit de l'anglais par Laurence Allard, Pierre Grosdemouge et Fred Pailler

Big data, la nécessité d'un débat

*Il nous a semblé intéressant de traduire, de façon collaborative (*via *Framapad), l'essai original que viennent de publier Danah Boyd et Kate Crawford présentant « Six provocations au sujet du phénomène des* big data *».*

Ces chercheuses, orientées vers l'ethnographie des usages des technologies de communication, s'interrogent – en toute connaissance de cause[1] *– sur les limites épistémologiques, méthodologiques, mais aussi éthiques des* big data *: champ d'études qui s'ouvre aujourd'hui sur la base des énormes jeux de données que fournit internet, en particulier celles générées par l'activité des usagers des sites de réseaux sociaux, que seuls des systèmes informatiques ont la capacité de collecter et de traiter.*

Les analyses des graphes relationnels de Facebook ou des flux de tweets *de Twitter sont des exemples bien connus de cette rencontre des sciences humaines et de l'informatique en réseau. Dans cet essai, les deux chercheuses personnifient ce champ de recherche en un* Big Data *faisant écho à Big Brother, et le confrontent à quelques principes méthodologiques des sciences humaines. Elles pointent également les dangers*

* L'article original a été présenté lors du Symposium sur les dynamiques de l'internet et de la société : « Une décennie avec internet », organisé par l'Oxford Internet Institute, le 21 septembre 2011, http://microsites.oii.ox.ac.uk/ics2011/content/home, consulté le 17 août 2012. La version originale de la traduction de cet article est disponible sur *InternetActu*, 23 septembre 2011, http://www.internetactu.net/2011/09/23/big-data-la-necessite-d%E2%80%99un-debat, consulté le 17 août 2012.

1. Voir cette étude sur les *tweets* des révolutions tunisienne et égyptienne à laquelle a participé Danah Boyd : Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, Danah Boyd, « The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions », *International Journal of Communication*, vol. 5, 2011, http://ijoc.org/ojs/index.php/ijoc/article/view/1246, consulté le 31 juillet 2012.

qu'une hégémonie mal anticipée de l'analyse automatisée des données risque de faire courir à la compréhension d'internet et de ses usages.

Répétons avec elles qu'un corpus n'est pas plus scientifique ou objectif parce que l'on est en mesure d'aspirer toutes les données d'un site. D'autant qu'il existe de nombreux biais (techniques avec les API, mais aussi organisationnels) dans l'accès même à ces données qu'on aurait tort de considérer comme totales. Cet accès ne repose en effet que sur le bon vouloir de sociétés commerciales et sur les moyens financiers dont disposent chercheurs et universités.

Ainsi, le phénomène des big data *risque, d'une part, de créer une nouvelle fracture numérique entre universités pauvres et riches, mais il peut aussi conduire à une vassalisation de la recherche scientifique par des sociétés commerciales et leurs services de marketing, utilisant eux aussi les* big data *pour profiler leurs produits.*

Ce virage computationnel des sciences humaines menace également de pérenniser inutilement le sempiternel clivage entre deux cultures scientifiques, l'une mathématique, objective par nature, et l'autre littéraire – subjective forcément. À moins qu'il ne soit vu comme une occasion de guérir enfin une partie des sciences humaines de leur péché interprétatif originel et de leur achiffrisme congénital.

Les studies *féministes, Donna Haraway*[2] *par exemple, ont démontré comment, du lexique jusqu'aux instruments d'observation utilisés, les pratiques scientifiques ne cessent d'être liées à la culture et à la société au sein desquelles elles sont pensables, et que leur travail d'explication chiffrée et de prévision des phénomènes naturels implique toujours une part d'interprétation. Les auteures soulèvent enfin les problèmes éthiques qu'implique l'usage des données personnelles des utilisateurs, lorsque celles-ci, bien que produites en public, sont loin d'être explicitement destinées aux usages scientifiques.*

Internet supporte aujourd'hui l'émergence d'une société de « citoyens-interprètes[3] *», c'est-à-dire potentiellement en capacité de produire et de traiter leurs propres données et connaissances dans les champs de la santé, de l'énergie ou encore de la politique. Cette diffusion des compétences interprétatives au sein de sociétés d'individus connectés, et l'accroissement de leurs capacités d'agir à partir des données qu'ils partagent volontairement, bref, la dimension profondément politique de ces activités en ligne, ne doivent pas se trouver noyés dans l'océan des* big data.

Ce texte suggère aussi que cette ère des big data *doit être accompagnée d'une réflexion politique au sein des* digital humanities. *Pour parodier Spiderman, avec Danah Boyd et Kate Crawford, n'oublions pas que* « With big power come big responsabilities ».

Laurence Allard, Pierre Grosdemouge, Fred Pailler

***

2. Donna Haraway, *Manifeste* cyborg *et autres essais. Sciences, fictions, féminismes*, Paris, Exils, 2007.

3. Yves Citton, *L'Avenir des humanités. Économie de la connaissance ou cultures de l'interprétation ?*, Paris, La Découverte, 2010.

> La technologie n'est ni bonne, ni mauvaise, ni neutre… L'interaction entre la technologie et l'écosystème social est telle que les développements techniques ont des conséquences environnementales, sociales, et humaines qui dépassent de loin les objectifs des appareils techniques et des pratiques elles-mêmes.
> Melvin Kranzberg[4]

> Nous devons ouvrir le débat – alors qu'il n'en existe aucun de sérieux actuellement – à propos des différentes temporalités, spatialités et matérialités que nous sommes susceptibles de représenter grâce à nos bases de données, avec, en vue, une conception permettant une flexibilité maximum, et autorisant, autant que possible, l'émergence d'une polyphonie et d'une polychronie. L'expression « données brutes » est un oxymore autant qu'une mauvaise idée ; au contraire, les données devraient être cuisinées avec soin.
> Geoffrey Bowker[5]

L'ère de *big data* a commencé. Les informaticiens, physiciens, économistes, mathématiciens, politologues, bio-informaticiens, sociologues, et beaucoup d'autres, réclament l'accès aux quantités massives d'informations produites par et à propos des gens, des choses, et de leurs interactions. Divers groupes discutent des coûts et des bénéfices de l'analyse de l'information issue de Twitter, Google, Verizon, 23andMe, Facebook, *Wikipedia*, et de tous les espaces dans lesquels de grands nombres de personnes laissent des traces numériques et déposent des données. D'importantes questions émergent. Les analyses de l'ADN à grande échelle aideront-elles à guérir les maladies ? Ou bien cela aboutira-t-il à une nouvelle vague d'inégalités médicales ? L'analyse des données rendra-t-elle l'accès des gens à l'information plus efficace et effectif ? Ou sera-t-elle plutôt utilisée pour pister les manifestants dans les rues des grandes villes ? Améliorera-t-elle la manière dont nous étudions la communication et la culture humaine, ou va-t-elle rétrécir la palette des options qui s'offrent à la recherche et altérer ce que « recherche » veut dire ? Tout ou partie de ces possibilités ?

Parler en termes de *big data* est, de bien des manières, restrictif. Comme l'observe Lev Manovitch[6], ce terme a été utilisé en sciences pour désigner les ensembles de données suffisamment grands pour nécessiter des superordi-

4. Melvin Kranzberg, « Technology and History: Kranzberg's Laws », *Technology and Culture*, vol. 27, n° 3, 1986, p. 545.

5. Geoffrey Bowker, *Memory Practices in the Sciences*, Cambridge, MIT Press, 2005, p. 183-184.

6. Lev Manovitch, « Trending: The Promises and the Challenges of Big Social Data » *in* M. K. Gold (ed.), *Debates in the Digital Humanities*, Minneapolis, The University of Minnesota Press, 2011, http://www.manovich.net/DOCS/Manovich_trending_paper.pdf, consulté le 15 juillet 2011.

nateurs, bien que, désormais, de grands ensembles de données puissent être analysés sur des ordinateurs de bureau avec des logiciels standards. Il n'y a aucun doute sur le fait que les quantités de données disponibles aujourd'hui soient en effet très grandes, mais ce n'est pas la caractéristique la plus pertinente de ce nouvel écosystème des données. Les *big data* sont remarquables, non en raison de leurs tailles, mais pour leurs capacités à être articulées à d'autres données. En raison des efforts pour exploiter et agréger les données, les *big data* sont fondamentalement liées aux réseaux. Leurs valeurs viennent des *patterns* qui peuvent être tirés du fait de connecter entre eux des jeux de données, concernant un individu, des individus liés à d'autres, des groupes de gens, ou simplement concernant la structure de l'information elle-même.

Plus encore, les *big data* sont importantes parce qu'elles renvoient à des analyses ayant cours à la fois à l'université et dans l'industrie. Au lieu de suggérer un terme nouveau, nous utilisons le terme *big data* ici en raison de sa prégnance populaire et parce que c'est le phénomène entourant les *big data* que nous souhaitons aborder. Ces *big data* amènent certains chercheurs à croire qu'ils peuvent tout voir d'une hauteur de 30 000 pieds. C'est le genre de données qui encourage la pratique de l'apophénie : voir des tendances là où il n'y en a aucune, simplement parce que des quantités massives de données peuvent offrir des connexions qui irradient dans toutes les directions. Pour cette raison, il est crucial de commencer à interroger les hypothèses qui vont gouverner l'analyse, les cadres méthodologiques, et les préjugés qui sous-tendent le phénomène *big data*.

Alors que les bases de données ont agrégé des données sur plus d'un siècle, le champ des *big data* n'est plus exclusivement le domaine des actuaires et des scientifiques. De nouvelles technologies ont rendu possible pour un grand nombre de personnes – incluant les chercheurs en humanités et en sciences sociales, les marketeurs, les organisations gouvernementales, les institutions éducatives, et les individus motivés – le fait de produire, partager, interagir avec, et organiser des données. Des jeux massifs de données autrefois illisibles et distincts, se trouvent articulés et aisément accessibles aujourd'hui. Les données deviennent chaque jour davantage notre « atmosphère numérique » : l'oxygène que nous inspirons et le dioxyde de carbone que nous expirons. Cet air est à la fois source de nourriture et de pollution.

La manière dont nous nous engageons dans l'ère des *big data* est cruciale : alors qu'elle s'installe dans un environnement d'incertitudes et de changements rapides, les décisions prises aujourd'hui auront un impact considérable dans le futur. Face à l'automatisation croissante de la collecte et de l'analyse des données – tels les algorithmes qui peuvent extraire et nous renseigner sur des *patterns* massifs dans le comportement humain –, il est nécessaire de se demander quels systèmes dirigent ces pratiques, et lesquels les régulent. Dans *Code*, L. Lessig[7] soutient que les systèmes sont régulés par

7. L. Lessig, *Code and Other Laws of Cyberspace*, 1999, http://code-is-law.org

quatre forces : le marché, la loi, les normes sociales, et l'architecture – ou, dans le cas de la technologie, le code.

Quand il s'agit des *big data*, ces quatre forces entrent en jeu, et, fréquemment, en conflit. Le marché voit les *big data* comme une pure opportunité : les marketeurs les utilisent pour orienter leurs campagnes, les assureurs veulent optimiser leurs offres, et les banquiers de Wall Street les utilisent pour améliorer leurs analyses des comportements du marché. Une législation a d'ores et déjà été proposée pour freiner la collecte et la rétention de données, généralement plutôt motivée par des questions de vie privée (par exemple, le *Do Not Track Online Act* de 2011 aux États-Unis). Des fonctionnalités comme la personnalisation permettent un accès rapide aux informations les plus pertinentes, mais elles entraînent de difficiles questions éthiques et divisent l'opinion de manière problématique[8].

Des études significatives et pertinentes qui s'appuient sur les méthodologies des *big data* sont actuellement réalisées, en particulier des études concernant les pratiques des sites de réseaux sociaux comme Facebook et Twitter. Néanmoins, il est impératif que nous commencions à poser des questions cruciales sur ce que signifient toutes ces données, qui y a accès, comment elles sont déployées, et à quelles fins. La montée des *big data* amène aussi de grandes responsabilités. Dans cet essai, nous proposons six provocations dont nous espérons qu'elles pourront éveiller les conversations sur les problèmes de *big data*. Il y a un enjeu pour les chercheurs du domaine des sciences sociales au cœur de la culture computationnelle du champ des *big data*, précisément dans la mesure où beaucoup de leurs questions centrales sont des questions fondamentales de nos disciplines. Aussi, nous croyons qu'il est temps de commencer à interroger de manière critique ce phénomène, ses hypothèses, ses partis pris.

L'automatisation de la recherche change la définition du savoir

Durant les premières décennies du XX^e^ siècle, Henry Ford a imaginé un système de production pour la fabrication de masse, utilisant des machines spécialisées et des produits standardisés. Sa vision est devenue la vision dominante du progrès technologique. Impliquant des chaînes d'automatisation et d'assemblage, le fordisme est devenu l'orthodoxie de la production pour les décennies suivantes : adieu les artisans compétents et le travail lent, bienvenue dans une ère du « fait à la machine[9] ». Mais il s'agissait de bien plus que d'un nouvel ensemble d'outils. Le XX^e^ siècle fut profondément

8. E. Pariser, *The Filter Bubble: What the Internet is Hiding from You*, New York, Penguin Press, 2011.

9. G. Baca, « Legends of Fordism: Between Myth, History, and Foregone Conclusions », *Social Analysis*, vol. 48, n° 3, 2004, p. 169-178.

marqué par le fordisme : ce dernier a produit une nouvelle compréhension du travail, de la relation humaine au travail et plus largement de la société.

Les *big data* ne renvoient pas uniquement aux très grands jeux de données et aux outils et procédures utilisés pour les manipuler et les analyser, mais aussi au tournant computationnel de la pensée et de la recherche[10]. Tout comme Ford a changé la manière dont nous fabriquons des voitures – et ainsi transformé le travail lui-même – les *big data* font émerger un système de savoir qui est déjà en train de transformer les objets de la connaissance, tout en ayant aussi le pouvoir d'informer la manière dont nous comprenons les réseaux humains et les communautés. « Changez les instruments, et vous changerez toute la théorie sociale qui va avec[11] », nous rappelle Latour.

Nous dirions que les *big data* créent un changement radical dans la manière dont nous pensons la recherche. Commentant la science sociale computationnelle, Lazer *et al.* affirment qu'elle offre « la capacité de collecter et d'analyser des données avec une ampleur, une profondeur et à une échelle sans précédent[12] ». Mais ce n'est pas qu'une question d'échelle. Pas plus qu'il ne suffit de considérer cela en termes de proximité, ou de ce que Moretti[13] évoque comme une analyse proche ou distante des textes. Il s'agit plutôt d'un profond changement au niveau de l'épistémologie et de l'éthique. Sont reformulées des questions clés concernant la constitution du savoir, le processus de recherche, la manière dont nous devons traiter l'information, la nature et la catégorisation de la réalité. Du Gay et Pryke ont noté que « les outils comptables [...] n'aident pas seulement à mesurer l'activité économique, ils donnent forme à la réalité qu'ils mesurent[14] ». De la même manière, les *big data* posent les bases de nouveaux terrains d'objets, de nouvelles méthodes de connaissance, de nouvelles définitions de la vie sociale.

Louant ce qu'il appelle « l'âge des Petabits », Chris Anderson, rédacteur en chef de *Wired*, écrit :

> C'est un monde dans lequel des quantités massives de données et les mathématiques appliquées remplacent tous les autres outils qui pourraient être utilisés. Exit toutes les théories sur les comportements humains, de la linguistique à la sociologie. Oubliez la taxinomie, l'ontologie, et la psycho-

10. L. Burkholder (ed.), *Philosophy and the Computer*, Boulder, Westview Press, 1992.

11. B. Latour, « Tarde's Idea of Quantification » *in* M. Candea (ed.), *The Social After Gabriel Tarde: Debates and Assessments*, London, Routledge, 2009, p. 9, http://www.bruno-latour.fr/articles/article/116-TARDE-CANDEA.pdf, consulté le 19 juin 2011.

12. D. Lazer *et al.*, « Computational Social Science », *Science*, vol. 323, 2009, p. 722.

13. F. Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History*, Londres, Verso, 2007.

14. P. du Gay, M. Pryke, *Cultural Economy: Cultural Analysis and Commercial Life*, London, Sage, 2002, p. 12-13.

> logie. Qui peut savoir pourquoi les gens font ce qu'ils font ? Le fait est qu'ils le font, et que nous pouvons le tracer et mesurer avec une fidélité sans précédent. Si l'on a assez de données, les chiffres parlent d'eux-mêmes[15].

Les chiffres parlent-ils d'eux-mêmes ? La réponse, pensons-nous, est un retentissant « NON ».

Le fait qu'Anderson congédie toutes les autres théories et disciplines est significatif : cela révèle l'existence d'un courant arrogant dans nombre de débats sur les *big data*, dans lesquels toutes les autres formes d'analyses peuvent être écartées au profit d'une production à la chaîne de chiffres, privilégiés comme étant en lien direct avec la connaissance brute. Les raisons pour lesquelles les gens font des choses, écrivent des choses, ou fabriquent des choses, sont effacées au profit du volume des répétitions numériques et de vastes modélisations. Ce n'est pas un lieu pour la réflexion, ni pour les formes plus anciennes d'habiletés intellectuelles. Comme David Berry l'écrit, les *big data* fournissent « des quantités déstabilisantes de connaissances et d'informations auxquelles il manque la force régulatrice de la philosophie[16] ». En lieu et place de la philosophie – que Kant voyait comme la base rationnelle de toute institution – « la computationalité pourrait alors être envisagée comme une ontothéologie, créant une nouvelle "époque" ontologique en tant que nouvelle constellation historique de l'intelligibilité[17] ».

Nous devons poser de difficiles questions sur les modèles d'intelligibilité des *big data* avant qu'elles ne se cristallisent en nouvelles orthodoxies. Si nous en revenons à Ford, son innovation utilisait la chaîne de montage pour fragmenter des tâches globales, interconnectées en tâches simples, atomisées et mécaniques. Il l'a fait en concevant des outils spécialisés qui prédéterminaient et limitaient fortement l'action du travailleur. De même, les outils spécialisés des *big data* intègrent leurs propres limitations et restrictions. L'une d'elles concerne le temps. « Les *big data* portent sur le présent exclusivement, sans le contexte historique qui est un facteur prédictif », observe Joi Ito, le directeur du MIT Media Lab[18]. Par exemple, Facebook et Twitter sont des sources de *big data* qui n'offrent que des fonctions limitées d'archivage et de recherche, et pour lesquelles les chercheurs auront tendance à

15. Chris Anderson, « The End of Theory. Will the Data Deluge Makes the Scientific Method Obsolete? », *Edge*, 2008, http://www.edge.org/3rd_culture/anderson08/%20anderson08_index.html, consulté le 25 juillet 2011. Au 31 juillet 2012, la page n'est plus accessible.

16. David Berry, « The Computational Turn: Thinking about the Digital Humanities », *Culture Machine*, vol. 12, 2011, p. 8 et 16, http://www.culturemachine.net/index.php/cm/article/view/440/470, consulté le 11 juillet 2011.

17. *Ibid.*, p. 12.

18. D. Bollier, *The Promise and Peril of Big Data*, Washington, The Aspen Institute, 2010, p. 19, http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf, consulté le 11 juillet 2011.

se concentrer sur des choses présentes ou sur le passé immédiat – traçant les réactions à une élection, une finale télévisée ou un désastre naturel – en raison de la difficulté même, voire de l'impossibilité, à accéder à des données plus anciennes.

Si nous observons l'automatisation de certains types particuliers de fonctions de recherche, alors nous devons considérer les défauts intégrés de ces machines-outils. Il ne suffit pas simplement de poser la question, comme le suggère Anderson, « Qu'est-ce que la science peut apprendre de Google ? », mais il faut se demander comment Google et les autres moissonneurs de *big data* peuvent changer le sens même d'apprendre, et quelles nouvelles possibilités et limites pourraient accompagner ces systèmes de connaissance.

Les revendications d'objectivité et d'exactitude sont trompeuses

« Des nombres, des nombres, des nombres », écrit Latour. « La sociologie a été obsédée par l'idée de devenir une science quantitative[19]. » Et pourtant, elle n'a toujours pas atteint ce but, selon Latour, puisqu'il dépend de l'endroit où l'on fait passer la ligne séparant la connaissance quantifiable de celle qui ne l'est pas en matière de social.

Les *big data* offrent aux humanités une nouvelle opportunité de revendiquer le statut de science quantitative aux méthodes objectives, en rendant quantifiables de plus en plus d'espaces sociaux. En réalité, travailler avec les *big data* reste une affaire subjective, et ce qui est quantifié ne peut pas forcément prétendre à une plus grande proximité avec une vérité objective – en particulier lorsque l'on considère les messages provenant des sites de médias sociaux. Pourtant, persiste la croyance erronée que les recherches qualitatives sont affaire d'interprétation de récits, et que les recherches quantitatives sont affaire de production de faits. Et c'est ainsi que les *big data* risquent de remettre à l'ordre du jour les divisions qui organisent les éternelles querelles sur les méthodes scientifiques.

La notion d'objectivité a constitué une question centrale pour la philosophie des sciences comme ce fut le cas lors des premiers débats sur les méthodes scientifiques[20]. D'un côté, la revendication de l'objectivité suggère une adhésion de la recherche à la sphère des objets, aux choses existant en elles-mêmes et pour elles-mêmes. D'un autre côté, la subjectivité est considérée avec suspicion, toute colorée qu'elle est par les diverses formes de conditionnements individuels et sociaux. La méthode scientifique s'efforce de se déprendre de toute subjectivité grâce l'application d'un processus dépassionné par lequel des hypothèses sont proposées et testées, aboutis-

19. B. Latour, *art. cit.*

20. É. Durkheim, 1895.

sant au final à une amélioration des connaissances. Néanmoins, les revendications d'objectivité sont nécessairement celles de sujets et sont fondées sur des observations et des choix subjectifs.

Tous les chercheurs sont des interprètes de données. Comme Lisa Gitelman[21] l'observe, les données doivent d'abord être imaginées, conçues comme des données, et ce processus d'imagination se base sur une forme d'interprétation : « chaque discipline institutionnalisée possède ses propres normes et standards concernant l'imagination des données ». Depuis que les chercheurs en informatique ont commencé à prendre part à la recherche en sciences sociales, il existe une tendance à considérer leurs travaux comme étant affaire de faits et non d'interprétations. Un modèle peut avoir l'air mathématiquement solide, une expérience peut sembler valide, mais dès lors que le chercheur tente d'en saisir le sens, le processus d'interprétation a commencé. Les décisions de conception, qui déterminent ce qui sera mesuré, découlent elles aussi d'un processus interprétatif.

Par exemple, dans le cas des données issues des médias sociaux, il existe un processus de « nettoyage des données » : des décisions sont prises pour savoir quels attributs et quelles variables vont être pris en compte, et lesquels vont être ignorés. Ce processus est intrinsèquement subjectif. Comme Bollier l'explique :

> En tant que grande masse de données brutes, les *big data* ne s'expliquent pas d'elles-mêmes. Qui plus est, les méthodologies spécifiques permettant d'interpréter les données sont soumises à toutes sortes de débats philosophiques. Les données peuvent-elles représenter une « vérité objective » ou bien est-ce que toute interprétation est forcément biaisée par une forme de filtrage subjectif, ou encore par la manière dont les données sont « nettoyées »[22] ?

Il faut ajouter à ces questions le problème des erreurs dans les données elles-mêmes. Les grands jeux de données récoltés sur internet sont souvent peu fiables, à la merci des pannes ou des pertes, et ces erreurs et lacunes sont décuplées quand on croise de multiples jeux de données. Les sociologues ont une longue histoire en termes de critique de la collecte des données et de vigilance à la façon dont un ensemble de biais peut influencer leurs données[23]. Une telle critique implique de comprendre les propriétés et les

21. Lisa Gitelman, « Notes for the Upcoming Collection "Raw Data" is an Oxymoron », 23 juillet 2011, https://files.nyu.edu/lg91/public

22. D. Bollier, *op. cit.*, p. 13.

23. M. Cain, J. Finch, « Towards a Rehabilitation of Data » *in* P. Abrams, R. Deem, J. Finch, P. Rock (eds.), *Practice and Progress: British Sociology 1950-1980*, London, George Allen and Unwin, 1981 ; J. Clifford, G. E. Marcus (eds.), *Writing Culture: The Poetics and Politics of Ethnography*, Berkeley, University of California Press, 1986.

limites d'un jeu de données, quelle que soit sa taille. Ce dernier peut contenir des millions et des millions de petits morceaux d'informations, mais cela ne signifie ni qu'il soit aléatoire ni qu'il soit représentatif. Pour avoir des prétentions statistiques face à un jeu de données, nous avons besoin de savoir d'où celles-ci proviennent ; et il est tout aussi important de connaître les faiblesses de ces données, et d'en rendre compte. Une telle démarche implique d'admettre que l'identité d'une personne et son point de vue informent les analyses qu'elle peut produire[24].

Des erreurs spectaculaires peuvent survenir lorsque les chercheurs tentent de faire des trouvailles en sciences sociales au sein des systèmes technologiques. Un exemple classique est né du choix de Friendster d'appliquer les travaux de Robin Dunbar[25]. Analysant la pratique du commérage chez les humains et de l'épouillage chez les singes, Dunbar trouva que les gens ne pouvaient entretenir activement plus de 150 relations, et défendait l'idée que ce nombre représentait la taille maximale du réseau personnel de quelqu'un. Malheureusement, Friendster a cru que les gens reproduiraient sur le site leur réseau personnel préexistant, et en a déduit que personne n'aurait une liste d'amis supérieure à 150. Il a donc bloqué le nombre d'« amis » que les gens pouvaient avoir sur ce service[26].

L'interprétation est au cœur de l'analyse de données. Quelle que puisse être la taille d'un jeu de données, il est sujet à des limitations et à des partis pris. Si ces limites et ces partis pris ne sont pas compris et soulignés, il faut s'attendre à des problèmes d'interprétation. Les *big data* atteignent le sommet de leur efficacité lorsque les chercheurs prennent en compte le processus méthodologique complexe qui sous-tend l'analyse de données sociales.

De plus grosses données ne sont pas toujours de meilleures données

Les chercheurs en sciences sociales ont longtemps affirmé que la rigueur de leur travail s'enracinait dans leur approche systématique de la collecte et de l'analyse de données[27]. Les ethnographes s'attachent à rendre compte réflexivement des préjugés que peuvent contenir leurs interprétations. Ceux

24. R. Behar, D. A. Gordon (eds.), *Women Writing Culture*, Berkeley, University of California Press, 1996.

25. Robin Dunbar, *Grooming, Gossip, and the Evolution of Language*, Cambridge, Harvard University Press, 1998.

26. Danah Boyd, « Friends, Friendsters, and Top 8: Writing Community into Being on Social Network Sites », *First Monday*, vol. 11, n° 12, article 2, 2006.

27. D. N. McCloskey, « From Methodology to Rhetoric » *in The Rhetoric of Economics*, Madison, University of Wisconsin Press, 1985, p. 20-35.

qui travaillent sur la base d'expérimentations contrôlent et standardisent la conception de leurs expériences. Les sociologues creusent la question des mécanismes de l'échantillonnage et des biais potentiellement contenus dans les questionnaires qu'ils utilisent dans leurs enquêtes. Les chercheurs quantitativistes soupèsent la représentativité statistique... Ce ne sont que quelques-unes des manières par lesquelles les chercheurs en sciences sociales essaient d'évaluer, chacun, la validité de leurs travaux respectifs. Malheureusement, certains de ceux qui abordent la question des *big data* supposent que ces questions au cœur des méthodologies des sciences sociales ne sont désormais plus pertinentes. On constate qu'un *ethos* sous-jacent pose ici problème, selon lequel plus gros signifie meilleur, quantité signifie nécessairement qualité.

Twitter fournit un bon exemple, dans le contexte d'une analyse statistique. Tout d'abord, Twitter ne représente pas « tout le monde », bien que beaucoup de journalistes et de chercheurs emploient « les gens » et « les usagers de Twitter » comme des synonymes. La population utilisatrice de Twitter n'est pas davantage représentative de la population globale. Et nous ne pouvons pas affirmer non plus qu'un compte Twitter équivaille à un utilisateur : certains utilisateurs ont plusieurs comptes, certains comptes sont utilisés par plusieurs personnes. Certaines personnes ne créent jamais de compte, mais accèdent à Twitter *via* le Web. Certains comptes sont en fait des « robots », qui produisent du contenu automatisé sans impliquer la présence d'une personne. Plus encore, la notion de compte « actif » est problématique. Tandis que certains usagers postent régulièrement du contenu sur Twitter, d'autres participent en tant que « écoutants[28] ». La société Twitter Inc. a révélé que 40 % des utilisateurs actifs ne se connectent que pour écouter[29]. Le sens véritable des termes « utilisateur », « participation » et « actif » doit donc être examiné de façon critique. En raison des incertitudes sur ce que représente véritablement un compte et sur les diverses formes que peut prendre l'engagement dans des activités liées au site, il serait aventureux de prendre un échantillon de comptes Twitter et d'en tirer des conclusions sur « les gens » ou « les utilisateurs ». Seul Twitter Inc. peut revendiquer un regard sur l'ensemble des comptes ou l'ensemble des *tweets* d'un échantillon aléatoire, dans la mesure où ils ont accès à la base de données centrale. Mais même ainsi, ils ne peuvent pas facilement comptabiliser les « voyeurs », ni les gens utilisant de multiples comptes ou les groupes de gens qui utilisent le même compte à plusieurs. Qui plus est, la base de données centrale est également sujette à des pannes, et les *tweets* sont fréquemment perdus et effacés.

28. K. Crawford, « Following You: Disciplines of Listening in Social Media », *Continuum: Journal of Media & Cultural Studies*, vol. 23, n° 4, 2009, p. 532.

29. Twitter, « One hundred million voices », *Twitter Blog*, 8 septembre 2011, http://blog.twitter.com/2011/09/one-hundred-million-voices.html, consulté le 12 septembre 2011.

Twitter Inc. rend accessible au public une fraction de son matériel, *via* ses API[30]. Le plus gros des flux offerts ainsi par Twitter, appelé par la firme elle-même le *firehose* (« la lance à incendie », *NdT*), permet d'accéder théoriquement à tous les *tweets* publics qui ont été postés et exclut explicitement tout *tweet* qu'un utilisateur aurait choisi de rendre privé ou « protégé ». Pourtant, certains *tweets* publiquement accessibles manquent encore dans le *firehose*. Bien qu'une poignée d'entreprises et de *start-up* puisse ainsi aspirer l'intégralité des *tweets*, très peu de chercheurs bénéficient d'un tel niveau d'accès. La plupart ont plutôt accès à ce que Twitter appelle le *gardenhose* (« le tuyau d'arrosage », *NdT*) (qui représente environ 10 % des *tweets* publics), ou même seulement au *spritzer* (« vin délayé », *NdT*) (environ 1 % des *tweets* publics), ou encore ont recours à une « liste blanche » de comptes grâce auxquels ils peuvent utiliser les API pour avoir accès à différents sous-ensembles de contenus tirés du flux public[31]. On manque donc d'informations permettant de savoir quels *tweets* sont exactement inclus dans ces différents flux de données et comment est construit leur échantillonnage. Il se peut que l'API n'extraie qu'un échantillon aléatoire de *tweets*, ou qu'elle ne retienne que les quelques premières centaines de *tweets* émis chaque heure, ou encore qu'elle ne retienne que les *tweets* issus d'un segment particulier du graphe du réseau. Étant donnée cette incertitude, il est difficile pour des chercheurs de revendiquer la qualité des données qu'ils sont en train d'analyser. Ces données sont-elles représentatives de tous les *tweets* ? Non, dans la mesure où elles excluent les *tweets* des comptes protégés[32]. Ces données sont-elles représentatives de tous les *tweets* publics ? Peut-être, mais pas nécessairement.

Ce ne sont là que quelques-unes des inconnues auxquelles les chercheurs font face lorsqu'ils travaillent sur les données de Twitter, pourtant ces limites sont rarement reconnues. Même ceux qui fournissent la procédure par laquelle ils ont construit leur échantillon à partir du *firehose* ou du *gardenhose* évoquent rarement ce qui pourrait manquer et comment leurs algorithmes ou l'architecture du système de Twitter peuvent introduire des distorsions dans le jeu de données. Certains chercheurs se concentrent simplement sur le nombre brut de *tweets*. Mais un grand nombre de données

30. API signifie *application programming interface* (*NdT* : interface de programmation) ; cela désigne un jeu d'outils que les développeurs utilisent pour accéder à des ensembles structurés de données.

31. Les détails des outils de développement fournis par Twitter peuvent être trouvés à l'adresse https://dev.twitter.com/docs/streaming-api/methods. Les comptes sur liste blanche constituaient au départ un mécanisme d'acquisition des autorisations d'accès, mais ils ne sont plus disponibles actuellement.

32. Le pourcentage de comptes protégés est inconnu. Dans une étude à travers laquelle ils ont tenté de repérer les comptes protégés et publics sur Twitter, Meeder *et al.* (« RT @IWantPrivacy: Widespread Violation of Privacy Settings in the Twitter Social Network », Communication lors de l'atelier « Web 2.0 Security and Privacy » (W2SP 2010), Oakland, 20 mai 2010) ont déterminé que 8,4 % des comptes identifiés étaient protégés.

(big data) et la totalité des données *(whole data)*, ce n'est pas la même chose. Si l'on ne peut prendre en compte le mode d'échantillonnage d'un jeu de données, sa taille n'est d'aucune importance.

Par exemple, un chercheur pourrait chercher à comprendre la fréquence de réactualisation des *tweets* en fonction des sujets abordés, mais si Twitter retire du flux tous les *tweets* qui contiennent certains mots ou certaines informations problématiques – des références à la pornographie par exemple – cette fréquence sera finalement complètement erronée. Indépendamment du nombre de *tweets*, un échantillon n'est pas représentatif si les données sont biaisées dès le départ. Twitter est devenu une source très populaire lorsqu'il s'agit d'exploiter des *big data*, mais travailler avec les données de Twitter pose de sérieux défis méthodologiques, rarement abordés par ceux qui s'y aventurent. Lorsque des chercheurs se mettent à travailler sur un jeu de données, ils ont besoin de comprendre – et de pouvoir expliquer publiquement – non seulement les limites de ce jeu de données, mais aussi les limites des questions qui peuvent se poser, et quelles sont les interprétations qui sont appropriées pour y répondre.

C'est particulièrement vrai lorsque les chercheurs combinent de multiples grands jeux de données. Jesper Anderson, le cofondateur du système de stockage de données financières ouvert FreeRisk, explique que le fait de combiner des données issues de multiples sources confronte à des défis particuliers : « Chacune de ces sources est sujette à des erreurs… Je pense que nous ne faisons qu'amplifier ce problème [quand on combine de multiples jeux de données][33]. » Cela ne signifie pas pour autant que combiner des données n'ait pas d'intérêt – certaines études, comme celle menée par Alessandro Acquisti et Ralph Gross[34], qui montrait comment les bases de données pouvaient être croisées pour révéler de très sérieuses violations de la vie privée, sont cruciales. Il est donc impératif que de telles combinaisons de données se fassent avec rigueur méthodologique et transparence.

Finalement, au tournant de l'ère computationnelle, il devient de plus en plus important de reconnaître la valeur des *« small data »*. Les intuitions de recherche peuvent apparaître à n'importe quel niveau, y compris à très petite échelle. Dans certains cas, se concentrer sur un seul individu peut s'avérer extraordinairement riche. On peut prendre pour exemple le travail de Tiffany Veinot[35], qui a suivi un seul travailleur – un inspecteur des voûtes dans une entreprise de services hydroélectriques – afin de comprendre les pratiques informationnelles des travailleurs en col bleu. En menant cette étude peu commune, Veinot a été amenée à déplacer la définition des

33. D. Bollier, *op. cit.*, p. 13.

34. Alessandro Acquisti, Ralph Gross, « Predicting Social Security Numbers from Public Data », *Proceedings of the National Academy of Science*, vol. 106, n° 27, 2009, p. 10975-10980.

35. Tiffany Veinot, « The Eyes of the Power Company: Workplace Information Practices of a Vault Inspector », *The Library Quarterly*, vol. 77, n° 2, 2007, p. 157-180.

« pratiques informationnelles », en s'écartant du regard porté habituellement sur leurs premiers usagers, les cols blancs, et en se rendant dans des espaces situés hors des contextes de l'entreprise ou de la ville. L'histoire que son travail nous raconte n'aurait pu être découverte en exploitant des millions de comptes Facebook ou Twitter, et si elle contribue de manière significative au champ de recherche, c'est en portant un regard sur le plus petit nombre possible de participants. La dimension des données reprises devrait ainsi correspondre à la question posée : dans certains cas, plus c'est petit, mieux c'est.

Toutes les données ne sont pas équivalentes

Certains chercheurs considèrent que les recherches menées sur de petits ensembles de données peuvent être améliorées grâce aux *big data*. Cet argument présuppose que les données sont interchangeables. Au contraire, sorties de leur contexte, les données perdent leur signification et leur valeur. Le contexte est déterminant. Si deux jeux de données peuvent être modélisés de la même manière, cela ne signifie pas pour autant qu'ils soient équivalents ni qu'ils puissent être analysés de la même façon. Considérons par exemple l'intérêt croissant pour l'analyse des réseaux sociaux qui a accompagné l'émergence des sites de réseaux sociaux[36] et l'obsession des industriels pour les « graphes sociaux ». Un nombre incalculable de chercheurs se sont rués sur Twitter et Facebook et sur d'autres médias sociaux pour analyser les graphes qui en résultaient, se découvrant des prétentions sur l'analyse des réseaux sociaux.

L'étude des réseaux sociaux date des débuts de la sociologie et de l'anthropologie[37], avec l'apparition de la notion de « réseau social » en 1954[38] et l'émergence du champ de l'« analyse des réseaux sociaux » peu de temps après[39]. Depuis lors, les universitaires de différentes disciplines ont tenté de comprendre les relations des gens entre eux en recourant à diverses approches méthodologiques et analytiques. Alors que les chercheurs commençaient à interroger les connexions entre les gens sur les médias sociaux en ligne, on a vu un véritable regain d'intérêt pour l'analyse des réseaux sociaux. Désormais, les spécialistes de l'analyse des réseaux se

36. D. Boyd, N. Ellison, « Social Network Sites: Definition, History, and Scholarship », *Journal of Computer-Mediated Communication*, vol. 13, nº 1, article 11, 2007.

37. Par exemple A. R. Radcliffe-Brown, « On Social Structure », *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, vol. 70, nº 1, 1940, p. 1-12.

38. J. A. Barnes, « Class and Committees in a Norwegian Island Parish », *Human Relations*, vol. 7, nº 1, 1954, p. 39-58.

39. L. Freeman, *The Development of Social Network Analysis*, Vancouver, Empirical Press, 2006.

tournent vers l'étude de ceux générés par les communications médiatisées, les déplacements géographiques et d'autres types de données traçables.

Cependant, les réseaux générés par les médias sociaux et résultant des traces communicationnelles ne sont pas nécessairement interchangeables avec les données issues des autres types de réseaux sociaux. Simplement parce que le fait que deux personnes soient physiquement coprésentes – ce qui pourrait être décelé par les antennes téléphoniques ou saisi par des photographies – ne signifie pas pour autant qu'elles se connaissent. En outre, plutôt que d'indiquer la présence de récurrences objectives et prévisibles, les sites de réseaux sociaux facilitent plutôt l'établissement de connexions qui traversent les frontières structurelles et agissent ainsi comme une source dynamique de changement : produire un instantané, ou même relever un ensemble de traces dans le temps, ne permet pas de saisir la complexité de toutes les relations sociales. Comme le notent Kilduff et Tsai, « les recherches sur les réseaux tendent à se baser sur une ontologie naïve qui considère comme non problématique l'existence et la persistance objectives de *patterns*, d'invariants et de systèmes sociaux[40] ». Cette approche produit un certain type de résultats lorsque l'analyse ne porte que sur un point déterminé dans le temps, mais elle s'effondre totalement dès lors que des questions plus vastes sont abordées[41].

Historiquement parlant, lorsque les sociologues et anthropologues s'intéressèrent, les premiers, aux réseaux sociaux, les données sur les relations entre individus étaient collectées par le biais d'enquêtes, d'entretiens, d'observations et de dispositifs d'expérimentation. Utilisant ces données, les sociologues se sont essentiellement attachés à décrire les « réseaux personnels » – l'ensemble de relations qu'un individu développe et entretient[42]. Ces connexions furent évaluées sur la base d'une série de mesures développées au fil du temps dans le but d'identifier les connexions personnelles. L'ère des *big data* introduit deux nouveaux types très populaires de réseaux sociaux, dérivés cette fois de l'étude des traces laissées par les données : les « réseaux articulés » et les « réseaux comportementaux ».

Les « réseaux articulés » sont ceux qui résultent du fait que les utilisateurs spécifient leurs contacts lorsqu'ils utilisent des technologies de médiation[43]. Il existe trois motifs fréquents pour lesquels les gens articulent ainsi leurs connexions : pour disposer d'une liste de leurs contacts à usage personnel,

40. M. Kilduff, W. Tsai, *Social Networks and Organizations*, Londres, Sage, 2003, p. 117.

41. D. Meyer *et al.*, « Organizing Far from Equilibrium: Nonlinear Change in Organizational Fields », *Organization Science*, vol. 16, n° 5, 2005, p. 456-473.

42. C. Fischer, *To Dwell Among Friends: Personal Networks in Town and City*, Chicago, University of Chicago, 1982.

43. Danah Boyd, « Friendster and Publicly Articulated Social Networks », Conférence « Human Factors and Computing Systems » (CHI 2004), ACM, Vienne, 24-29 avril 2004, http://www.danah.org/papers/CHI2004Friendster.pdf, consulté le 31 juillet 2012.

pour afficher publiquement leurs connexions à certains autres, et pour filtrer le contenu sur les médias sociaux. On trouve ces réseaux articulés sous la forme de carnets d'adresses mails ou téléphoniques, de listes de contacts de messageries instantanées, de listes d'« amis » sur certains réseaux sociaux, et de « *followers* » sur d'autres types de réseaux sociaux. Les motivations qui poussent les gens à ajouter quelqu'un à chacune de ces listes sont très variables, mais le résultat reste que ces listes peuvent inclure des amis, des collègues, des connaissances, des célébrités, des personnalités publiques, et des inconnus jugés intéressants.

Les « réseaux comportementaux » sont dérivés de l'analyse des modes de communication, des coordonnées téléphoniques et des interactions sur les médias sociaux[44]. Ils peuvent inclure les personnes qui s'envoient des SMS, celles qui sont taguées ensemble sur des photos sur Facebook, les gens qui s'envoient des e-mails, et les gens qui se trouvent physiquement dans les mêmes espaces, du moins si l'on se fie à ce qu'indiquent leurs téléphones portables.

Réseaux « articulés » et « comportementaux » ont tous deux une grande valeur aux yeux des chercheurs, mais ils ne sont pas équivalents aux réseaux personnels. Par exemple, bien que souvent contesté, le concept de « force des liens » est conçu pour indiquer l'importance des relations individuelles[45]. Quand une personne choisit de lister quelqu'un parmi ses « meilleurs amis » sur Myspace, il peut s'agir véritablement, ou pas, d'un de ses amis les plus proches ; il existe toutes sortes de raisons sociales de ne pas mentionner ses plus intimes connexions au sommet de la liste[46]. De même, lorsque les téléphones mobiles permettent de repérer qu'un travailleur passe plus de temps avec ses collègues qu'avec son épouse, cela ne signifie pas pour autant qu'il entretient des liens plus forts avec ses collègues qu'avec sa femme. Mesurer la force des liens au seul prisme de leur fréquence ou des articulations publiques est une erreur courante : la notion de force des liens – et de bien des théories qui se sont construites autour – exige une estimation subtile de la manière dont les gens envisagent et valorisent leurs relations avec les autres.

De fascinantes analyses de réseaux peuvent être réalisées à partir de ces réseaux articulés et comportementaux. Mais il existe un risque, à l'ère des *big data*, de traiter chaque connexion comme équivalente à toutes les autres, de confondre la fréquence des contacts avec la force des relations, et de croire

44. M. R. Meiss *et al.*, « Structural Analysis of Behavioral Networks from the Internet », *Journal of Physics A: Mathematical and Theoretical*, vol. 41, n° 22, 2008, p. 220-224 ; J. P. Onnela *et al.*, « Structure and Tie Strengths in Mobile Communication Networks », *Proceedings from the National Academy of Sciences*, vol. 104, n° 18, 2007, p. 7332-7336.

45. M. S. Granovetter, « The Strength of Weak Ties », *American Journal of Sociology*, vol. 78, n° 6, 1973, p. 1360-1380.

46. Danah Boyd, « Friends, Friendsters, and Top 8: Writing Community into Being on Social Network Sites », *cit.*

qu'une absence de connexion indique qu'une relation devrait être établie. Les données ne sont pas génériques. Il y a certes un intérêt à analyser des données abstraites, mais le contexte demeure crucial.

Accessible ne veut pas dire éthique

En 2006, un projet de recherche basé à Harvard a commencé par rassembler les profils de 1700 étudiants usagers de Facebook afin d'étudier comment leurs centres d'intérêts et leurs amitiés évoluaient avec le temps[47]. Ces données prétendument anonymes ont été rendues accessibles à tous, permettant à d'autres chercheurs de les explorer et de les analyser. Ces autres chercheurs ont, en revanche, rapidement découvert qu'il était possible de désanonymiser certaines parties de ce jeu de données, compromettant ainsi la vie privée des étudiants, dont aucun ne savait que ces données avaient été collectées[48]. Cette affaire fit les gros titres des journaux, et posa un problème épineux aux universitaires : quel statut accorder à des données dites « publiques » sur les réseaux sociaux ? Peuvent-elles être simplement utilisées, sans en demander la permission ? Quelle serait la démarche la plus éthique pour les chercheurs ? Les militants pour la protection de la vie privée y voient d'ores et déjà un champ de bataille crucial, sur lequel l'établissement de meilleurs dispositifs de protection de la vie privée s'avère nécessaire. Toute la difficulté réside dans le fait que les brèches dans la vie privée sont délicates à spécifier – peut-on en constater les dégâts au moment même où elles ont lieu ? Et qu'en sera-t-il vingt ans après ? « Tout type de donnée portant sur des sujets humains soulève des questions de protection de la vie privée, et il est difficile de quantifier les véritables risques induits par l'usage abusif de ces données[49]. »

Même lorsque les chercheurs s'efforcent de procéder avec précaution, ils ne sont pas toujours conscients des dommages que leurs recherches pourraient entraîner. Par exemple, un groupe de chercheurs avait noté qu'il existait une corrélation entre le fait de s'automutiler (le « *cutting* ») et le suicide. Ils avaient préparé une intervention pédagogique cherchant à décourager les gens de s'automutiler ainsi, pour finir par apprendre que cette intervention induisait une augmentation des tentatives de suicide. Pour certains,

47. K. Lewis *et al.*, « Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.com », *Social Networks*, vol. 30, 2008, p. 330-342.

48. M. Zimmer, « More on the "Anonymity" of the Facebook Dataset–It's Harvard College », *Michael Zimmer.org*, 2008, http://michaelzimmer.org/2008/01/03/more-on-the-anonymity-of-the-facebook-dataset-its-harvard-college, consulté le 20 juin 2011. Au 31 juillet 2012, la page n'est plus accessible.

49. *Nature*, cité *in* David Berry, *art. cit.*

en effet, les automutilations servaient de soupape de sécurité et tenaient à distance le désir de se suicider. Les scientifiques cessèrent immédiatement leurs interventions[50].

Les comités d'éthique dédiés à la recherche sont apparus dans les années soixante-dix pour superviser la recherche sur l'humain. Bien que leur mise en œuvre ait incontestablement été problématique[51], le but de ces comités est de fournir un cadre permettant d'évaluer les dimensions éthiques de certaines recherches par enquêtes, et de s'assurer que de bons contrepoids sont mis en place pour protéger les personnes. Des pratiques comme le « consentement éclairé » et la protection de la vie privée des informateurs sont destinées à donner du pouvoir aux participants, compte tenu des abus qui ont pu avoir lieu au sein des sciences médicales et sociales[52]. Bien que les comités d'éthique ne soient pas toujours en mesure de prévoir les méfaits d'une étude en particulier – et viennent, trop souvent, empêcher les chercheurs de se lancer dans des recherches pour des motifs autres qu'éthiques –, l'intérêt de l'existence de ces comités reste d'inciter les universitaires à une pensée critique quant à l'éthique de leurs recherches.

Alors que les *big data* commencent à émerger en tant que champ de recherche, on comprend encore bien peu de choses quant aux implications éthiques des recherches mises en œuvre. Sur quelles bases inclure quelqu'un dans un vaste ensemble de données ? Que se passe-t-il si un billet « public » sur le blog de quelqu'un est sorti de tout contexte et analysé d'une manière que son auteur n'aurait jamais imaginée ? Que signifie pour quelqu'un le fait d'être mis sous les projecteurs ou d'être « étudié » sans même le savoir ? Qui est responsable de s'assurer qu'un processus de recherche ne s'avère pas nuisible pour des individus ou des communautés ? Que devient le consentement ?

Il ne serait pas raisonnable de demander aux chercheurs d'obtenir le consentement de chacune des personnes qui poste un *tweet*, mais il n'est pas éthique, de leur part, de légitimer leurs actions par le simple fait que les données sont accessibles[53]. La déontologie de la collecte et de l'analyse

50. T. Emmens, A. Phippen, *Evaluating Online Safety Programs*, Cambridge, Harvard Berkman Center for Internet and Society, 2010, http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Emmens_Phippen_Evaluating-Online-Safety-Programs_2010.pdf, consulté le 23 juillet 2011.

51. Z. M. Schrag, *Ethical Imperialism: Institutional Review Boards and the Social Sciences, 1965-2009*, Baltimore, Johns Hopkins University Press, 2010.

52. T. Blass, *The Man Who Shocked the World: The Life and Legacy of Stanley Milgram*, New York, Basic Books, 2004 ; S. M. Reverby, *Examining Tuskegee: The Infamous Syphilis Study and Its Legacy*, Chapel Hill, University of North Carolina Press, 2009.

53. D. Boyd, A. Marwick, « Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies », Communication lors du symposium « Une décennie avec internet » organisé par l'Oxford Internet Institute, Oxford, 22 septembre 2011.

des données en ligne révèle de très sérieuses problématiques[54]. Le processus d'évaluation éthique de la recherche ne peut pas être simplement ignoré parce que les données sont apparemment accessibles. Les chercheurs doivent continuer à s'interroger – et à interroger leurs collègues – sur la déontologie de leurs collectes de données, de leurs analyses, et de leurs publications.

S'ils souhaitent agir de manière éthique, il est important que les universitaires réfléchissent à l'importance de leur responsabilité. Dans le cas des *big data*, cela renvoie à la fois à une responsabilité devant le champ de recherche et à une responsabilité devant les sujets de la recherche. Lorsqu'ils travaillent avec des participants humains, les chercheurs académiques sont tenus au respect de standards professionnels spécifiques afin que soient protégés leurs droits et leur bien-être. Néanmoins, le problème est que beaucoup d'instances de supervision éthique ne comprennent pas les processus d'exploitation et d'anonymisation des *big data*, sans parler des erreurs qui peuvent rendre les données personnelles identifiables. La responsabilité devant le champ et devant les sujets humains requiert une pensée rigoureuse de toutes les ramifications des *big data*, plutôt que la seule supposition que les comités d'éthique vont nécessairement faire ce qu'il faut pour s'assurer que les gens soient protégés. La responsabilité est ici utilisée dans un sens plus large que la simple protection de la vie privée, comme Troshynski *et al.*[55] l'ont souligné, dans la mesure où le concept de responsabilité peut s'appliquer même lorsque les attentes conventionnelles en termes de vie privée ne sont pas remises en cause. La responsabilité renvoie ici davantage à une relation multidirectionnelle : il peut y avoir responsabilité devant des supérieurs, des collègues, des participants, et devant l'opinion publique[56].

Les études de *big data* recèlent d'importantes questions sur la vérité, le contrôle et le pouvoir : les chercheurs disposent des outils et des accès, tandis que les utilisateurs des médias sociaux, dans leur ensemble, n'en disposent pas. Leurs données ont été produites dans des espaces dont le contexte s'avère particulièrement sensible et déterminant, et il est fort probable que certains utilisateurs de médias sociaux n'accorderaient pas leur permission pour que leurs données soient utilisées ailleurs. Beaucoup n'ont pas conscience de la multiplicité d'agents et d'algorithmes qui collectent et stockent leurs données pour des usages ultérieurs. Les chercheurs sont rarement le public qu'un utilisateur s'imagine avoir, pas plus que les utilisateurs ne sont nécessairement conscients des multiples usages, profits et autres bénéfices qui peuvent être tirés des informations qu'ils ont mises en ligne. Les données peuvent être publiques (ou semi-publiques), mais cela ne doit

54. C. Ess, « *Raw Data* » *is an Oxymoron*, 2002, https://files.nyu.edu/lg91/public/Data.pdf, consulté le 23 juillet 2011.

55. Troshynski *et al.*, 2008.

56. Dourish & Bell, 2011.

pas être pris, de façon simpliste, comme une permission totale, donnée pour toute forme d'utilisation. Il existe une différence considérable entre le fait d'être en public et celui d'être public, différence qui est rarement reconnue par les chercheurs du champ des *big data*.

L'accès limité aux *big data* crée de nouvelles fractures numériques

Dans un essai sur les *big data*, Scott Golder[57] cite le sociologue Georges Homans[58] : « Les méthodes des sciences sociales sont coûteuses en temps et en argent et deviennent plus coûteuses encore chaque jour. » Historiquement, la collecte de données a effectivement toujours été difficile, chronophage et coûteuse. L'essentiel de l'enthousiasme autour des *big data* provient de l'impression qu'elles offrent au contraire un accès facile à un grand nombre de données.

Mais qui y a accès ? Avec quels objectifs ? Dans quels contextes ? Et avec quelles contraintes ? Bien que l'explosion de la recherche utilisant des jeux de données tirés des médias sociaux donne à croire que l'accès est devenu simple et direct, c'est tout sauf vrai. Comme Lev Manovich le fait remarquer,

> seules les entreprises de médias sociaux ont accès à des bases de données sociales véritablement conséquentes – et plus particulièrement aux données concernant les interactions et les échanges. Un anthropologue travaillant pour Facebook ou bien un sociologue travaillant chez Google accéderont à des données auxquelles le reste de la communauté scientifique n'accédera jamais[59].

Certaines entreprises empêchent complètement l'accès à leurs données. D'autres vendent à bon prix ce privilège de l'accès. Et d'autres encore cèdent de petits jeux de données aux chercheurs travaillant pour des universités. Tout ceci produit des écarts de niveaux considérables dans le système de la recherche : ceux qui ont des moyens financiers – ou bien ceux qui travaillent au sein des entreprises – peuvent conduire des recherches très différentes de ceux qui sont dehors. Ceux qui n'ont accès à rien ne peuvent ni reproduire ni donc évaluer les affirmations méthodologiques de ceux qui bénéficient d'un accès privilégié.

57. Scott Golder, « Scaling Social Science with Hadoop », *Cloudera Blog*, 5 avril 2010, http://www.cloudera.com/blog/2010/04/scaling-social-science-with-hadoop, consulté le 18 juin 2011.

58. Georges Homans, *Social Behavior: Its Elementary Forms*, Cambridge, Harvard University Press, 1974.

59. Lev Manovich, *art. cit.*

Il est également important de reconnaître que la classe des « riches » des *big data* se trouve renforcée par le système universitaire : les universités les mieux cotées et les mieux dotées sont les seules capables d'acheter l'accès aux données, et les étudiants des grandes universités sont les plus susceptibles d'être invités à travailler pour les grandes entreprises de médias sociaux. Ceux qui restent en périphérie se verront moins probablement proposer ces invitations et trouveront donc moins l'occasion de développer leurs compétences. Il en résulte que l'écart entre ceux qui sont allés dans les universités prestigieuses et les autres se creusera significativement.

Au-delà des questions d'accès, il y a également des questions de compétences. Batailler avec les API, fouiller et analyser de grands pans de données est une compétence généralement réservée aux personnes expérimentées en informatique. Lorsque les compétences informatiques deviennent les plus valorisées, émerge la question de savoir qui se trouve avantagé ou désavantagé par un tel contexte. Cela crée de nouvelles hiérarchies tournant autour de « qui saura lire les chiffres », plutôt que la reconnaissance qu'informaticiens et sociologues peuvent offrir chacun des points de vue valables. De façon significative, il s'agit également d'une différence entre les genres. La plupart des chercheurs qui ont des compétences en informatique aujourd'hui sont des hommes, et, comme les historiens féministes et les philosophes des sciences l'ont montré, l'identité de celui qui pose les questions détermine les questions qui seront posées[60]. C'est un point difficile que d'évaluer le type de compétences de recherche qui sera valorisé dans le futur, et la manière dont ces compétences seront enseignées. Que faut-il enseigner aux étudiants pour qu'ils soient aussi à l'aise avec les algorithmes et l'analyse de données qu'avec l'analyse sociologique et la théorie ?

En définitive, la difficulté et le coût de l'accès aux données des *big data* aboutissent à une culture étriquée des résultats de recherche. Les grandes entreprises de données n'ont aucune obligation de rendre leurs données disponibles, et ont un contrôle total sur le choix de ceux qui y accèdent. Les chercheurs du champ des *big data* qui ont accès à ces jeux de données propriétaires sont moins susceptibles de choisir des questions qui pourraient être litigieuses pour une société de médias sociaux, par exemple, s'ils pensent que cela peut aboutir à l'interruption de leur droit d'accès. Les effets dissuasifs sur les types de questions de recherche qui peuvent être posés – en public comme en privé – sont une chose dont nous devons tous tenir compte pour évaluer l'avenir des *big data*.

L'écosystème qui entoure actuellement les *big data* crée un nouveau type de fracture numérique : des *big data* de riches et des *big data* de pauvres.

60. D. Forsythe, *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*, Stanford, Stanford University Press, 2001 ; S. Harding, « Feminism, Science and the anti-Enlightenment Critiques » *in* L. J. Nicholson (ed.), *Feminism/Postmodernism*, New York, Routledge, 1990.

Les chercheurs de certaines grandes entreprises sont même allés jusqu'à suggérer que les universitaires ne devraient pas venir entraver l'étude des médias sociaux – les « chercheurs-maison » pouvant s'en occuper tellement plus efficacement[61]. De tels efforts pour distinguer des chercheurs initiés de chercheurs étrangers et profanes – ce qui n'a rien de nouveau – mettent à mal la rhétorique utopiste des évangélistes des valeurs des *big data*. « Une démocratisation effective peut toujours se mesurer à ce critère essentiel », affirmait Derrida, « la participation et l'accès aux archives, à leur constitution et à leur interprétation[62] ». Chaque fois que les inégalités sont explicitement inscrites au sein même d'un système, elles produisent des structures qui reconduisent des différences de classe. Manovich décrit trois classes d'individus au royaume des *big data* : « ceux qui créent les données (que ce soit consciemment ou en laissant des traces numériques), ceux qui ont les moyens de les recueillir, et ceux qui ont la compétence de les analyser[63] ». Nous savons que ce dernier groupe est le plus restreint, mais aussi le plus privilégié : c'est également celui qui arrive à déterminer les règles selon lesquelles les *big data* seront exploitées, et à choisir qui pourra participer à cette exploitation. Bien que les inégalités institutionnelles puissent parfois être considérées comme inéluctables par le monde universitaire, elles doivent néanmoins être examinées et interrogées, dans la mesure où elles orientent les données comme les types de recherches susceptibles d'en émerger.

Affirmer que le phénomène des *big data* participe de certains des plus grands changements historiques et philosophiques ne revient pas à suggérer qu'il en soit le seul responsable. Le monde académique n'est en aucun cas l'unique moteur du tournant computationnel. Il existe un mouvement de fond, gouvernemental et industriel, pour récolter et extraire le maximum de valeur des données, qu'il s'agisse d'informations qui permettront de rendre plus efficaces les publicités, le design de produits, la planification du trafic ou la lutte contre le crime. Mais nous croyons réellement qu'il existe de nombreuses et sérieuses conséquences à l'opérationnalisation des *big data*, et à ce que cela va signifier pour l'agenda scientifique. Comme Lucy Suchman[64] l'observe, *via* Lévi-Strauss, « nous sommes nos outils ».

61. Durant son discours à la Conférence internationale sur les blogs et les médias sociaux (ICWSM), à Barcelone, le 19 juillet 2011, Jimmy Lin – chercheur travaillant chez Twitter – décourageait les chercheurs de se lancer dans des projets de recherche pouvant être menés à bien plus facilement par les chercheurs travaillant chez Twitter, compte tenu de leur accès privilégié aux données de Twitter.

62. Jacques Derrida, *Archive Fever: A Freudian Impression*, trad. Eric Prenowitz, Chicago, University of Chicago Press, 1996, p. 4.

63. Lev Manovich, *art. cit.*

64. Lucy Suchman, « Consuming Anthropology » *in* Andrew Barry, Georgina Born (eds.), *Interdisciplinarity: Reconfigurations of the Social and Natural Sciences*, Londres/New York, Routledge, 2011.

Lorsque nous les utilisons, nous devrions donc également prendre en considération la manière dont ils participent à la construction du monde. L'ère des *big data* vient à peine de commencer, mais il est d'ores et déjà important que nous nous mettions à interroger les hypothèses, les valeurs, et les partis pris de cette nouvelle vague de recherches. En tant qu'universitaires investies dans la production de la connaissance, de telles interrogations constituent une part essentielle de ce que nous faisons.

Remerciements des auteures

Nous voulons remercier Heather Casteel pour son aide dans la préparation de cet article. Nous sommes aussi profondément reconnaissantes envers Eytan Adar, Tarleton Gillespie et Christian Sandvig pour leurs conversations inspirantes, leurs suggestions et leurs retours sur ce texte.

Remerciements des traducteurs

Merci à Samuel Ripault et Laëtitia Tin pour leur aide précieuse.