D'abord les données, ensuite la méthode?

Big data et déterminisme en sciences sociales

Plantin, J.-C.; Russo, F.

DOI
10.4000/socio.2328

Publication date
2016

Document Version
Final published version

Published in
Socio

Socio

La nouvelle revue des sciences sociales

6 | 2016 :
Déterminismes
Dossier : Déterminismes
Droit de suite

D’abord les données, ensuite la méthode ?

Big data et déterminisme en sciences sociales*

First the data, then the method? Big data and determinism in the social sciences

Jean-Christophe Plantin et Federica Russo

p. 97-115

Résumés

Français English
Si les chercheurs en sciences sociales ont depuis longtemps recours à de larges quantités de données, par exemple avec les enquêtes par questionnaire, le recours à des données numériques massives et hétérogènes, ou « *big data* », est de plus en plus fréquent. À travers un abandon de la théorie pour la recherche de corrélations, cette multitude de données suscite-t-elle une nouvelle forme de déterminisme ? L’histoire des sciences sociales indique au contraire que l’accroissement des données disponibles a entraîné un rejet progressif d’une hypothèse déterministe héritée des sciences de la nature, au profit d’une autonomisation méthodologique fondée sur la modélisation statistique. Dans ce contexte, cet article montre que l’accent mis sur la taille des *big data* ne signifie pas tant un retour au déterminisme, mais est davantage révélateur du désajustement actuel entre les caractéristiques de ces données massives et les méthodes et infrastructures en sciences sociales.

While large quantities of data have long been used by social science researchers, for example survey questionnaires, the use of massive and heterogeneous digital data, or “big data” is more and more frequent. As theory is abandoned in the search for correlations, is this multitude of data promoting a new form of determinism? On the contrary, the history of the social sciences demonstrates that the increase in data available has led to a gradual phasing out of a determinist hypothesis inherited from the natural sciences, to the benefit of a methodological empowerment based on statistical modelling. In this context, this article demonstrates that the emphasis on the size of the big data is not necessarily indicative of a return to determinism; instead it tends to reveal the present mismatch between the characteristics of these massive quantities of data and the methods and infrastructures in the social sciences.

Entrées d'index

Mots-clés : big data, démographie, déterminisme, infrastructure, probabilité, sciences sociales
Keywords : big data, demography, determinism, infrastructure, probability, social sciences

Notes de la rédaction
**Socio* a publié dans son no 4 un dossier intitulé « Le tournant numérique… et après » auquel ce droit de suite fait écho.

Texte intégral

1 Le terme de « *big data* », désormais de plus en plus employé dans les sciences sociales, prend son origine dans le secteur de l'économie numérique et se place à la suite de notions telles que « Web 2.0 » ou « Web de plateformes ». Dans ce contexte, il renvoie à la fois à un objet et à un processus. En tant qu'objet, le terme désigne des données de grande taille (d'où « *big* » *data*), mais également stratégiques et autrefois conservées en silos de données internes ou non exploitées. Celles-ci proviennent de sources hétérogènes et non structurées : autant de propriétés souvent résumées à travers trois V : volume, variété et vélocité (Dumbill, 2012). En tant que processus, le terme désigne les différents systèmes d'analyse et de stockage pour traiter et valoriser ces données, avec un accent mis sur les nouvelles capacités de traitement en fonction de cette taille (Magoulas et Lorica, 2009)[1].

2 Au-delà de ce secteur originel, le « déluge de données » concerne également les sciences dites parfois « dures », fait attesté par les numéros des revues *Nature* en 2008 et *Science* en 2011 consacrés à ce sujet. Les *big data* prennent ainsi place au sein d'une multitude de disciplines, par exemple les sciences informatiques, les sciences de l'ingénieur, les télécommunications, ou les mathématiques (Halevi et Moed, 2012), ainsi qu'en sciences biomédicales (Costa, 2014). La masse des données en question appelle également la mise en place de larges infrastructures pour soutenir ces recherches. L'infrastructure de données en astronomie dans le cadre du projet Sloan Digital Sky Survey[2], ou le « défi Mastodons » du Centre national de la recherche scientifique[3] constituent autant d'exemples d'engagements pluridisciplinaires dans le développement de capacités pour la recherche à partir de ces grandes masses de données.

3 Les sciences humaines et sociales ne sont toutefois pas en reste. Ainsi, les « sciences sociales computationnelles » (Lazer *et al.*, 2009) proposent l'analyse de grands ensembles de données numériques (provenant par exemple de courriels ou de téléphones mobiles) à travers les outils et méthodes de l'analyse de réseaux (Barabási et Albert, 1999 ; Watts et Strogatz, 1998), afin d'étudier comment les individus interagissent, se déplacent ou communiquent. Également, même si ce n'est pas nécessairement leur objectif premier, les travaux du champ des humanités numériques peuvent porter sur une masse importante de données (Plantin et Monnoyer-Smith, 2014). Les travaux en « humanités computationnelles » (Manovich, 2012) prennent ainsi en compte la totalité des œuvres d'un peintre ou la collection exhaustive des couvertures du magazine *Time*. Si les conséquences des *big data* pour les sciences humaines, et principalement les humanités numériques, ont déjà été au cœur d'un dossier spécial de la revue *Socio* (2015), notre article interroge exclusivement les changements et continuités que les données massives apportent aux sciences sociales.

4 Quelle est la relation entre ces masses de données et la notion de déterminisme en sciences sociales ? On entend par déterminisme soit l'idée que la réalité (tant physique que biologique, ou sociale) est régie par des lois (déterministes) et qu'il n'*existe* donc pas de phénomènes aléatoires ; soit l'idée que si, nous, agents

épistémiques, *connaissions* toutes les lois (de la nature, de la société, etc.) et toutes les conditions initiales, nous pourrions *prédire* avec certitude tout état futur du monde. Pierre-Simon de Laplace, dans son *Essai philosophique sur les probabilités* (1814), imagine un « Démon » omniscient qui pourrait tout prédire avec exactitude. Mais cela n'est qu'une idéalisation de notre connaissance imparfaite : pour nous, agents épistémiques « limités » et « imparfaits », il ne reste qu'à employer le calcul de probabilité pour réaliser des prévisions. Dans cette optique, l'usage répandu des modèles probabilistes, tant en sciences naturelles qu'en sciences sociales, ne serait pas dû à la nature intrinsèquement aléatoire des phénomènes, mais à la nature incertaine de la connaissance, qui affaiblit la possibilité d'établir des lois. Au XIXe siècle, les sciences sociales ont fondé la bataille pour leur légitimation en tant que sciences (au même titre que les sciences naturelles, en particulier la physique) sur la possibilité de découvrir des lois régissant la sphère sociale. Notre discussion dans cet article se limite au déterminisme comme certitude de nos connaissances et de nos prédictions dans le contexte des sciences sociales.

5 Selon les néopositivistes, les hypothèses, venant des théories, guident la récolte et l'analyse des données. Toutefois, les bases de données disponibles de plus en plus considérables semblent renverser cette démarche et nous faire revenir à la tradition empiriste : la formulation des hypothèses et des théories est dérivée de l'exploration des données. Si cette idée est déjà fortement critiquée en sciences sociales, il est toutefois important de l'interroger à travers l'histoire de la discipline : les *big data* contemporaines suscitent-elles le retour d'une conception déterministe des sciences sociales ?

6 Une perspective historique sur les différents afflux de données en sciences sociales permet d'étudier cette question à la lumière des conséquences de cet accroissement de données sur l'interaction entre données, théories et hypothèses. Cette analyse des relations précédentes, à partir de la seconde moitié du XIXe siècle, entre masse de données et appareillage théorique des sciences sociales fournit des points de comparaison avec les *big data* contemporaines, et conduit à contester l'hypothèse d'un retour du déterminisme.

7 Comme nous allons le voir, l'autonomisation des sciences sociales concernant le statut de l'explication a au contraire entraîné un abandon progressif du modèle déterministe – provenant initialement de la forte influence des sciences de la nature –, afin de prendre en compte la variabilité intrinsèque de l'objet de recherche. L'exemple des études de la population en sciences sociales permet d'illustrer en quoi l'afflux de données disponibles a précipité l'émergence d'un modèle probabiliste, et a renforcé cet abandon d'un modèle déterministe du monde social. Nous montrerons enfin que si les *big data* contemporaines semblent renouveler la tentation déterministe en sciences sociales, ceci reflète davantage une inadéquation entre les données disponibles et les moyens pour les traiter : ces données ne paraissent alors « massives » que parce qu'elles n'ont pas encore suscité un renouvellement des modalités de leur traitement et de leur classification.

L'autonomisation de l'explication en sciences sociales

8 L'arrivée des *big data* en sciences sociales revitalise-t-elle une tendance au déterminisme, c'est-à-dire l'idée qu'on puisse donner des explications complètes de tout phénomène et des prévisions précises sur l'évolution de la société ? En effet, l'acquisition massive d'information à l'aide des *big data* pourrait suggérer que l'on se rapproche d'une science laplacienne, où tout est connu et prévisible. Avant d'analyser les différentes prises de position actuelles sur cette question, il est important de rappeler que cet espoir déterministe n'est pas nouveau en sciences sociales. La

société, à travers ses principes, ses mécanismes et son évolution, a toujours été au centre de la réflexion philosophique et scientifique : il suffit de renvoyer à la pensée grecque sur l'économie – au sens de la gestion de l'administration du foyer – ou sur la *politeia* – au sens du discours sur la gouvernance et la citoyenneté. Toutefois, nous voulons mettre ici en évidence un épisode beaucoup plus récent dans la méthodologie des sciences sociales, à savoir l'introduction de la statistique au XIXe siècle, afin d'illustrer cette tension entre espoir déterministe et déluge de données[4]. Adolphe Quetelet, physicien et astronome de formation, applique de manière systématique les statistiques en sciences sociales. Dans son ouvrage *Essai de physique sociale* (1835), son projet est d'étudier la société à l'image de la physique, afin de trouver les lois (déterministes) qui régissent l'évolution de l'« homme moyen ». Ce dernier est une idéalisation de l'« homme réel », étudié selon certaines caractéristiques, à l'image de l'étude du mouvement en physique où les objets sont modélisés sans masse et sans dimension. Quetelet met les sciences sociales sur la piste d'une méthodologie « objective », qui ouvre les portes à d'autres contributions importantes, telles que la sociologie quantitative d'Émile Durkheim (1895). Les sciences sociales actuelles disposent dès lors de méthodologies quantitatives sophistiquées et puissantes, notamment grâce à l'élan donné par des méthodologues comme Quetelet ou Durkheim.

9 Toutefois, cet élan déterministe s'est très vite heurté à une difficulté répandue, et, selon certains auteurs, intrinsèque aux sciences sociales, à savoir la variabilité de l'objet d'investigation. Les êtres humains relèvent en effet d'une hétérogénéité en termes de cultures qui empêche d'atteindre le niveau de généralisation des sciences naturelles. Mais les sciences sociales se heurtent également à d'autres sources de variabilité, telles que la variabilité psychologique. Les êtres humains perturbent encore cette objectivité prétendue pour deux raisons supplémentaires. D'une part, prenant connaissance des descriptions de leur comportement, ils sont susceptibles de modifier ce dernier en conséquence (ce problème est aussi appelé « performativité »). D'autre part, les chercheurs mêmes peuvent interférer et changer le comportement des personnes qu'ils étudient (ce problème est aussi appelé « réflexivité »). Dès ce tournant objectiviste, il est clair que les sciences sociales ne pourront atteindre l'« objectivité » de la physique.

10 Les sciences sociales et la philosophie qui s'y intéresse ont toutefois fait un long chemin pour affirmer leurs propres critères de scientificité. Ce très long débat au sein des champs de la sociologie et de la philosophie des sciences a atteint son sommet lors de la « guerre des sciences » (Ashman et Baringer 2001 ; Latour, 2001 : chap. VII). La nature de la science même est alors en jeu, à travers deux conceptions opposées : d'une part, le point de vue orthodoxe, selon lequel la science est objective, rationnelle et dénuée de valeurs, et d'autre part le point de vue constructiviste et postmoderne, selon lequel la science est profondément déterminée par des facteurs socio-politiques qui la rendent hautement contextuelle et contingente. Il a fallu, à la suite de ce débat, passer par une réévaluation du statut des objets étudiés par les sciences sociales et de la notion d'objectivité (Montuschi, 2003, 2004, 2008 ; Daston et Galison, 2012), tout en admettant l'impossibilité d'aboutir à des explications et à des prévisions comme la physique. Le débat a également mis l'accent sur la rigueur de la méthode, plutôt que sur le caractère déterministe des modèles ou sur l'existence objective des objets étudiés[5].

11 Le statut de l'explication et de la preuve en sciences sociales est donc passé de la tentation déterministe, directement calquée sur les sciences de la nature, à une prise en compte de la variabilité intrinsèque aux sociétés analysées. Les méthodes statistiques et probabilistes ont joué un rôle important à cet égard.

Les études de la population et

l'explication déterministe

12 L'intérêt pour une étude quantitative de la population remonte aux XVIIIe et XIXe siècles : il suffit de penser aux travaux de Euler (1760), Malthus (1798) ou Ravenstein (1885, 1889), qui étudiaient notamment la croissance et les migrations des populations. L'approche dominante était alors celle des modèles déterministes et mécanistes provenant de la physique. Adophe Quetelet voulait utiliser les méthodes de ces disciplines pour étudier les causes du développement de l'« homme moyen ». Par conséquent, établir la (ou les) cause(s) d'un phénomène permettrait aussi d'en saisir l'évolution et d'en faire des prévisions précises. Il ressort de ces démarches un point de vue causaliste très explicite, notamment à travers une vision où le monde (tant naturel que social) est régi par des lois déterministes. Dans cette optique, un lien étroit entre déterminisme et causalité est décelable, qui a dominé le XXe siècle, remis en question cependant au cours des cinquante dernières années par la philosophie de la causalité. Pour une fois, la philosophie n'a pas été la « chouette de Minerve », comme le pensait Hegel, elle a au contraire suivi la cadence des sciences, qui ont, les premières, embrassé le tournant probabiliste au détriment d'une vue déterministe[6].

13 L'émergence des études quantitatives de la population a précipité cet abandon du déterminisme par le biais de l'usage massif des statistiques. Lorsque la taille des populations étudiées s'est accrue et que, par conséquent, les bases de données se sont élargies, il a fallu avoir recours à d'autres outils statistiques, notamment l'échantillonnage, et dans un temps ultérieur, les logiciels statistiques (Courgeau, 2012). Si, d'une part, cela a permis d'étudier des populations de plus en plus grandes, cela a, d'autre part, introduit une composante aléatoire dans l'analyse et l'explication des phénomènes sociaux. Les techniques d'échantillonnage présentent l'avantage de permettre de faire des inférences sur une population *entière* en n'étudiant qu'une partie. Mais, en même temps, nous devons prévoir la possibilité de nous tromper dans ces inférences, car elles ne se fondent que sur une *partie* de la population. Cela montre le caractère inductif, et donc faillible, de l'inférence statistique. Ce premier déluge de données en sciences sociales a ainsi pour conséquence l'abandon du modèle déterministe en faveur d'une vision stochastique de la réalité sociale.

14 Sans doute le développement du traitement automatique de données, puis l'informatique ont aidé considérablement le traitement et l'analyse de grandes masses de données, tout en renforçant la nécessité de l'échantillonnage. Par exemple, l'automatisation du traitement des données de la population commence dès la fin du XIXe siècle. Herman Hollerith, le futur créateur d'IBM, développe à cette période le système de tabulation mécanique (Hollerith Electric Tabulating System), qui remporta la compétition pour le traitement des données du recensement de 1890 aux États-Unis (Campbell-Kelly, 1990 : 124) : cette méthode, reconnue comme la plus efficace, devient la technologie officielle et se trouve dès lors largement utilisée dans l'administration.

15 Nous espérons avoir montré, à l'aide de la discussion sur les études quantitatives de la population, qu'un accroissement de la quantité de données et une sophistication des moyens pour les traiter n'alimentent pas directement un modèle déterministe. Au contraire, l'accroissement des données disponibles faisant émerger la nécessité de l'échantillonnage et de l'automatisation du calcul statistique, ce sont davantage les modèles probabilistes qui se développent, pour constituer la base des méthodes contemporaines quantitatives en sciences sociales (aussi bien, par exemple, en démographie qu'en économie).

Big data et sciences sociales : un

retour du déterminisme ?

16 Nous avons vu dans les deux parties précédentes que les sciences sociales n'ont cessé de se différencier du modèle déterministe des sciences naturelles pour développer leur propre modèle d'explication et leurs concepts. Toutefois, les *big data* ne remettent-elles pas en cause cet effort ? L'afflux récent de données numériques massives dans plusieurs disciplines scientifiques suscite actuellement de nombreux débats méthodologiques et épistémologiques parmi la communauté scientifique[7]. À la suite de González-Bailón (2013), il est possible d'identifier plusieurs enthousiastes quant aux potentiels de ces nouvelles sources de données pour la recherche scientifique. Ainsi, cet afflux de données est présenté comme un changement épistémologique majeur, un « quatrième paradigme » où la découverte scientifique intensive en données prendrait la suite des paradigmes scientifiques précédents (fondés successivement sur l'observation, la théorie et la simulation, voir Hey *et al.*, 2009). Le journaliste Chris Anderson (2008) a formulé une version grand public de cette hypothèse à travers son célèbre article paru dans le magazine *Wired* : « The data deluge makes the scientific method obsolete ». Anderson y défend l'idée d'une obsolescence des modèles d'analyse face à la masse et à la granularité des données disponibles (recueillies par des dispositifs de collecte en temps réel, non obstrusifs et proches des pratiques quotidiennes), et des traitements computationnels pour les analyser. Avec suffisamment de données, les nombres parleraient ainsi d'eux-mêmes.

17 Ce positionnement épistémologique a des conséquences méthodologiques propres. Mayer-Schönberger et Cukier citent trois possibilités de recherche qui émergent des *big data* :

> The first is the ability to analyze vast amounts of data about a topic rather than be forced to settle for smaller sets. The second is a willingness to embrace data's real-world messiness rather than privilege exactitude. The third is a growing respect for correlations rather than a continuing quest for elusive causality[8] (Mayer-Schönberger et Cukier, 2013 : 29).

18 À travers ces trois points, les auteurs soulignent les avantages d'une démarche d'analyse exploratoire, partant des données et formulant des questions de recherche après avoir exploré celles-ci (Janowicz *et al.,* 2014). La découverte de corrélations à travers cette démarche serait alors plus efficace qu'une méthode traditionnelle, partant d'une hypothèse à tester et visant à découvrir une causalité dans les phénomènes étudiés. Il est important de remarquer que les propos de Mayer-Schönberger et Cukier font ressortir une notion plutôt ancienne de la causalité, fortement liée au déterminisme. Leur troisième point, en particulier, semble opposer corrélation et causalité, comme si cette dernière, qui selon leurs termes est « insaisissable », ne se trouverait que dans des lois déterministes. Mais la philosophie de la causalité a depuis évolué en fonction des apports de la modélisation statistique et probabiliste (Russo, 2009). Au cours des dernières décennies, la philosophie a développé plusieurs concepts adaptés aux défis contemporains des sciences : par exemple, le concept de « causalité probabiliste » n'apparaît plus comme un oxymore, il est tout à fait adapté à la modélisation probabiliste tant en sciences sociales qu'en sciences naturelles. Cependant, ce tournant probabiliste ne nous force pas à rejeter la formule selon laquelle « la corrélation n'est pas la causalité ». Bien au contraire, cela a poussé tant les philosophes que les scientifiques à expliciter les conditions sous lesquelles les corrélations constituent une preuve en faveur de l'existence des relations causales[9].

19 À rebours de ces points de vue enthousiastes, les implications de ces masses de données pour la recherche en sciences sociales sont actuellement au centre de nombreux débats[10]. Certains auteurs réfutent l'hypothèse précédente d'un inversement de l'appareil hypothético-déductif, au profit d'une pratique scientifique

guidée par l'exploration de données et mettant au second plan l'appareillage théorique. Ainsi, c'est le mythe du commencement de la recherche par l'exploration de données qui est contesté. L'identification et l'extraction de données – processus souvent présentés comme objectifs – résultent déjà d'un choix parmi les données disponibles afin de différencier celles jugées plus importantes pour l'analyse (Drucker, 2011). Oublier ou passer sous silence cette présélection fait ainsi courir le risque d'une réification des données. D'autres auteurs soulignent que les données brutes relèvent de l'oxymore (Gitelman, 2013), car les données sont toujours déjà une représentation construite du phénomène analysé (Kurgan, 2013)[11]. Une rigueur scientifique et un effort de réflexivité appellent à la mise en lumière de ces choix qui prennent place *avant* l'étape des données, par exemple à travers une analyse des outils de visualisation (Drucker, 2011) ou des systèmes de classification adoptés (Bowker et Star, 1999). Enfin, la prééminence de la grande taille des données dans ces recherches est également à interroger : il a été montré en sciences du climat (Edwards, 2010) et en épidémiologie (Ioannidis, 2008) qu'un accroissement des données peut se révéler contre-productif. Pour le premier cas d'étude, la modélisation et la simulation se révèlent plus efficaces qu'un accroissement des données disponibles ; pour le second, cela peut aller jusqu'à alimenter une crise de confiance envers la validité des résultats scientifiques (Plantin *et al.*, à paraître). Ainsi, les *big data* font l'objet de nombreuses critiques en sciences sociales, résumées par Kitchin (2014) : elles ne sont pas exhaustives, ne proviennent pas de nulle part, elles ne sont pas exemptes de théories, et ne parlent pas d'elles-mêmes.

Désajustement entre données et infrastructures

20 Tout accroissement de données rend celles-ci « massives » au premier abord, jusqu'à ce qu'une mutation des systèmes de traitement les rende utilisables pour la recherche. Les sciences n'en sont pas à leur premier épisode d'accroissement massif de données : Strasser rappelle ainsi que « *perceptions of an "information overload" (or a "data deluge") have emerged repeatedly from the Renaissance through the early modern and modern periods and each time specific technologies were invented to deal with the perceived overload*[12]. » (Strasser, 2012 : 85). Des cas précédents dans l'histoire des sciences l'illustrent, celui des savants de la Renaissance qui durent faire face à un afflux de données en provenance du Nouveau Monde, ou encore, celui de la montée en puissance des collectionneurs à l'époque des Lumières qui s'est traduite par un afflux massif de spécimens aux muséums d'histoire naturelle, bien au-delà de ce que ces institutions pouvaient étudier. « La production et l'application de standards » (*ibid.* : 86) s'avèrent alors nécessaires, et prennent notamment la forme de nouveaux systèmes de classification, tels que celui créé par Carl von Linné (Wright, 2007 : 132) au XVIIIe siècle. Ainsi, bien au-delà d'un changement de paradigme scientifique, les mutations portent essentiellement sur les systèmes de traitement visant à rendre les données accessibles et manipulables pour la recherche.

21 Cette inadéquation entre *big data* actuelles et modalités de leur traitement est formulée par Lagoze (2014) comme une sortie hors de la « zone de contrôle » (Atkinson, 1996) des données. Cette notion désigne originellement la bibliothèque, avec toute sa chaîne d'acteurs, de systèmes de classification et de bâtiments garantissant l'intégrité, la disponibilité et la stabilité à long terme des documents (Lagoze, 2014). Dès lors, quelle est la zone de contrôle traditionnelle en sciences sociales, et en quoi les *big data* en constituent-elles une sortie ?

22 Les sciences sociales se sont en grande partie structurées au lendemain de la Seconde Guerre mondiale autour de la recherche quantitative fondée sur le partage

de données. Dans le cas des États-Unis, les grandes enquêtes par questionnaire ont commencé dès le recensement de 1890, mais il faudra attendre plus de cinquante ans pour que de larges études par questionnaires et statistiques gouvernementales constituent la base de travaux en sciences sociales (on citera ainsi le *Roper Poll,* le *General Social Survey,* l'*American National Election Studies* et le *Current Population Survey* [Converse, 2009]). Afin d'accompagner l'utilisation et la dissémination de ces nouvelles sources de données, plusieurs infrastructures de recherche, également appelées cyberinfrastructures (Jackson *et al.*, 2007) ont été créées à cette époque : l'Inter-University Consortium for Political and Social Research (ICPSR) en 1962, Essex Archive en 1967, Norwegian Social Science Data en 1971, et le Consortium of European Social Science Data Archives (CESSDA) en 1976. Au-delà de leurs spécificités, ces institutions constituent autant de zones de contrôle pour les données en sciences sociales : leur rôle est de garantir toutes les étapes de la chaîne de provenance des données (Lagoze, 2014) afin d'assurer leur fiabilité lors de leur réutilisation et leur archivage. L'ICPSR à l'université du Michigan vise par exemple à acquérir, archiver et disséminer les données des chercheurs afin de favoriser leur réutilisation. Cet objectif passe par un travail de préparation des données, au cours duquel la structuration des fichiers de données est vérifiée, les données incorrectes sont modifiées, la documentation (notamment sur le codage adopté) est créée et les métadonnées de l'étude sont insérées au catalogue.

23 Toutefois, les *big data* présentent des caractéristiques techniques et légales qui ne correspondent pas au fonctionnement des infrastructures de recherche en sciences sociales, et rendent ainsi difficile leur prise en charge par celles-ci. La comparaison des propriétés des larges fichiers de données provenant d'une part du service Twitter et d'autre part de l'ICPSR met en avant trois points de divergence entre *big data* et infrastructures. Tout d'abord, le service Twitter applique des mesures relatives à la propriété intellectuelle pour ses données qui empêchent leur reproduction et leur large dissémination (Beurskens, 2013). Précisément, les termes d'utilisation (« *terms of service* ») actuels de Twitter comportent une clause interdisant la redistribution, la vente et la location des données[13]. Cette politique de non-republication a pour conséquence majeure que les chercheurs ayant recours à des fichiers de données provenant de Twitter (principalement acquis à travers son interface de programmation, « Application Programming Interface » [Gaffney et Puschmann, 2013]) ne peuvent pas republier ces données en ligne : cette contrainte compromet la réplicabilité des résultats de ces études, mais entrave également leur prise en charge par les infrastructures de recherche, qui par définition visent à la redistribution de ces données.

24 De plus, les infrastructures de recherche citées ci-dessus ont développé avec les années leur propre dépendance (ou « *path dependence* ») à l'égard des enquêtes par questionnaires en sciences sociales : toutes les capacités humaines et techniques ont ainsi été orientées vers ce type de données. Pour l'ICPSR, les logiciels utilisés pour le traitement des données relèvent des logiciels d'analyse statistique, tel le Statistical Package for the Social Sciences (SPSS). De même, le travail des processeurs de données passe par l'utilisation de différents scripts visant à faciliter le travail de « nettoyage » des données (par exemple identifier les valeurs manquantes, ou les codes inconsistants) : ces scripts se fondent sur la structure des fichiers d'enquêtes par questionnaire. Ainsi, les compétences requises pour travailler au sein de cette infrastructure s'appuient sur des méthodes traditionnelles en sciences sociales, qui se révèlent inadaptées pour prendre en compte les caractéristiques techniques, légales et méthodologiques des *big data*.

25 L'exemple de l'usage scientifique des données de Twitter illustre bien l'inadéquation entre les propriétés actuelles des *big data* et le fonctionnement des infrastructures de recherche en sciences sociales. Les *big data* requièrent autre chose que les capacités traditionnelles de traitement des larges fichiers de données en

sciences sociales. Les données n'en paraissent donc que davantage *massives*, du fait que les méthodes existant pour leur dissémination ne permettent pas de les rendre accessibles pour la recherche.

Conclusion

26 Cet article a interrogé les conséquences épistémologiques de l'accroissement des données numériques disponibles pour la recherche en sciences sociales et battu en brèche un hypothétique retour au déterminisme. L'histoire des sciences sociales a connu plusieurs épisodes d'accroissement de données, conduisant à l'abandon de l'hypothèse déterministe, et à une autonomisation des concepts en sciences sociales. Chaque apparition de données en grand nombre s'est traduite par l'émergence et l'adaptation des systèmes de classification, d'archivage et de dissémination de ces données ; si les *big data* apparaissent ainsi « massives », c'est parce que leurs caractéristiques diffèrent des données traditionnellement utilisées en sciences sociales, et nécessitent un ajustement en termes de capacités de traitement et de dissémination, encore à venir.

27 Au-delà du débat sur les capacités des infrastructures de recherche, l'émergence des *big data* soulève la question de la compatibilité des buts du secteur privé avec ceux de la recherche. En effet, l'entrée de ces compagnies du Web dans le domaine scientifique peut avoir des conséquences importantes sur la relation entre chercheurs, sources de données et méthodes. Tout d'abord, le développement de la figure du chercheur invité (*in-house researcher*). Les chercheurs invités bénéficient alors d'un accès à de très larges contenus, voire à la totalité des données de l'entreprise, tout en étant également soumis aux contraintes de la propriété industrielle et de l'anonymisation des données. Récemment, plusieurs entreprises du Web ont également participé aux modes de valorisation de la communauté scientifique : Facebook a organisé en 2014 une préconférence en direction des chercheurs en sociologie, un jour avant le congrès annuel de l'American Sociological Association[14] ; Twitter a lancé la même année un appel à projets scientifiques (les *Twitter Data Grants*), gratifiant les projets retenus d'un accès préférentiel aux données.

28 Tous ces éléments ont pour conséquence des glissements potentiellement conflictuels entre démarche industrielle et recherche scientifique. Le récent scandale public autour de l'étude des sentiments sur Facebook (Kramer, Guillory et Hancock, 2014)[15] montre un décalage entre les potentiels pour la recherche (avoir accès à 689 003 sujets d'étude) et les précautions légales et méthodologiques de protection des sujets[16]. Ici encore, la taille des *big data* ne constitue qu'un des défis que représentent ces données pour la recherche en sciences sociales : un grand nombre de débats à venir porte davantage sur l'aspect éthique de ces recherches.

Bibliographie

Anderson, Chris, 2008, « The end of theory: The data deluge makes the scientific method obsolete », *WIRED*, 23 juin. Consultable en ligne : <http://www.wired.com/science/discoveries/magazine/16-07/pb_theory>.

Ashman, Keith M. et Baringer, Philip S. (éd.), 2001, *After the Science Wars*, Londres et New York, Routledge.

Atkinson, Ross, 1996, « Library functions, scholarly communication, and the foundation of the digital library: Laying claim to the control zone », *The Library Quarterly*, vol. 66, no 3, p. 239-265.
DOI : 10.1086/602884

BARABÁSI, Albert-László et ALBERT, Reka, 1999, « Emergence of scaling in random networks », *Science,* vol. 286, nº 5439, p. 509-512.

BEURSKENS, Michael, 2013, « Legal questions of Twitter research », dans Katrin Weller, Axel Bruns, Jean Burgess et Merja Mahrt, *Twitter and Society*, New York, Peter Lang International Academic Publishers, p. 123-136.

BOGEN, James et WOODWARD, James, 1988, « Saving the phenomena », *The Philosophical Review*, vol. XCVIII, nº 3, juillet, p. 303-352.
DOI : 10.2307/2185445

BOWKER, Geoffrey C. et STAR, Susan Leigh, 1999, *Sorting Things Out: Classification and Its Consequences*, Cambridge, The MIT Press.

BOYD, danah et CRAWFORD, Kate, 2012, « Critical questions for big data », *Information, Communication & Society,* vol. 15, nº 5, p. 662-679.

CAMPBELL-KELLY, Martin, 1990, « Punch-card machinery », dans William Aspray (éd.), *Computing Before Computers*, Ames, Iowa State University Press, p. 122-155.

COASE, Ronald H., 1988, « How should economists choose? », dans *Ideas, Their Origins and Their Consequences: Lectures to Commemorate the Life and Work of G. Warren Nutter,* Washington D.C., American Enterprise Institute for Public Policy Research.

CONVERSE, Jean M., 2009, *Survey Research in the United States: Roots and Emergence 1890-1960*, New Brunswick, Transaction Publishers.

COSTA, Fabricio F., 2014, « Big data in biomedicine », *Drug Discovery Today*, vol. 19, nº 4, p. 433-440.
DOI : 10.1016/j.drudis.2013.10.012

COURGEAU, Daniel, 2012, *Probability and Social Science,* Dordrecht, Springer.
DOI : 10.1007/978-94-007-2879-0

CRAWFORD, Kate, 2014, « The anxieties of big data », *The New Inquiry*, 30 mai. Consultable en ligne : <http://thenewinquiry.com/essays/the-anxieties-of-big-data/>.

DASTON, Lorraine et GALISON, Peter, 2012, *Objectivité*, traduit de l'anglais par Sophie Renaut et Hélène Quiniou, Dijon, Presses du réel.

DRUCKER, Johanna, 2011, « Humanities approaches to graphical display », *Digital Humanities Quarterly*, vol. 5, nº 1. Consultable en ligne : <http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html>.

DUMBILL, Edd, 2012, « What is big data? An introduction to the big data landscape », *O'Reilly Radar*, 11 janvier. Consultable en ligne : <http://radar.oreilly.com/2012/01/what-is-big-data.html>.

DURKHEIM, Émile, 1967 [1895], *Les règles de la méthode sociologique*, Paris, Presses universitaires de France.

EDWARDS, Paul N., 2010, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming,* Cambridge, The MIT Press.

EULER, Leonhard, 1760, « Recherches générales sur la mortalité et la multiplication du genre humain », *Histoire de l'Académie royale des sciences et des belles lettres de Berlin*, nº 16, p. 144-164.

First Monday, 2013, numéro spécial : « Making data – Big data and beyond », vol. 18, nº 10, octobre.

GAFFNEY, Devin et PUSCHMANN, Cornelius, 2013, « Data collection on Twitter », dans Katrin Weller, Axel Bruns, Jean Burgess et Merja Mahrt, *Twitter and Society*, New York, Peter Lang, p. 55-68.

GITELMAN, Lisa, 2013, « *Raw Data » Is an Oxymoron*, Cambridge, The MIT Press.

GONZÁLEZ-BAILÓN, Sandra, 2013, « Social science in the era of big data », *SSRN : Social Science Research Network*, mars. Consultable en ligne : <http://papers.ssrn.com/abstract=2238198>.
DOI : 10.1002/1944-2866.POI328

HALEVI, Gali et MOED, Henk F., 2012, « The evolution of big data as a research and scientific topic: Overview of the literature », *Research Trends*, nº 30, septembre. Consultable en ligne : <http://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/>.

HEY, Tony, TOLLE, Kristin et TANSLEY, Stewart, 2009, *The Fourth Paradigm Data-Intensive Scientific Discovery*, Microsoft Research.

ILLARI, Phyllis et RUSSO, Federica, 2014, *Causality: Philosophical Theory Meets Scientific Practice*, Oxford, Oxford University Press.

International Journal of Communication, 2014, numéro special : « Critiquing big data:

Politics, ethics, epistemology », nº 8.

IOANNIDIS, John P. A., 2008, « Why most discovered true associations are inflated », *Epidemiology*, vol. 19, nº 5, p. 640-648.
DOI : 10.1097/EDE.0b013e3181813le7

JACKSON, Steven J., EDWARDS, Paul N., BOWKER, Geoffrey C. et KNOBEL, Cory P., 2007, « Understanding infrastructure: History, heuristics and cyberinfrastructure policy », *First Monday,* vol. 12, nº 6.
DOI : 10.5210/fm.v12i6.1904

JANOWICZ, Krzysztof, VAN HARMELEN, Frank, HENDLER, James et HITZLER, Pascal, 2014, « Why the data train needs semantic rails », *AI Magazine,* 1er janvier. Consultable en ligne : <http://corescholar.libraries.wright.edu/cse/169>.

Journal of Broadcasting and Electronic Media, 2013, dossier spécial : « Emerging methods for digital media research », vol. 57, nº 1.

KITCHIN, Rob, 2014, « Big data, new epistemologies and paradigm shifts », *Big Data & Society*, vol. 1, nº 1.
DOI : 10.1177/2053951714528481

KRAMER, Adam D. I., GUILLORY, Jamie E. et HANCOCK, Jeffrey T., 2014, « Experimental evidence of massive-scale emotional contagion through social networks », *Proceedings of the National Academy of Sciences,* vol. 111, nº 24, p. 8788-8790.
DOI : 10.1073/pnas.1320040111

KURGAN, Laura, 2013, *Close up at a Distance: Mapping, Technology, and Politics,* New York, Zone Books.

LAGOZE, Carl, 2014, « Big data, data integrity, and the fracturing of the control zone », *Big Data & Society*, novembre. Consultable en ligne : <http://bds.sagepub.com/content/1/2/2053951714558281>.
DOI : 10.1177/2053951714558281

LAPLACE, Pierre-Simon de, 1814, *Essai philosophique sur les probabilités*, Paris, Vve Courcier.
DOI : 10.1259/jrs.1921.0077

LATOUR, Bruno, 2001, *L'espoir de Pandore. Pour une version réaliste de l'activité scientifique*, Paris, La Découverte.

LAZER, David, PENTLAND, Alex, ADAMIC, Lada, ARAL, Sinan, BARABÁSI, Albert-László, BREWER, Devon, CHRISTAKIS, Nicholas *et al.,* 2009, « Life in the network: The coming age of computational social science », *Science,* vol. 323, nº 5915, p. 721-723.

MAGOULAS, Roger et LORICA, Ben, 2009, « Big data: Technologies and techniques for large-scale data », dans Jimmy Guterman, *Release 2.0: Issue 11*, Sebastopol, O'Reilly.

MALTHUS, Thomas Robert, 1798, *An Essay on the Principles of Population,* Londres, J. Johnson, in St. Paul's Church-yard.

MANOVICH, Lev, 2012, « Software studies: Computational humanities vs. digital humanities », *Software Studies Initiative,* 16 mars. Consultable en ligne : <http://lab.softwarestudies.com/2012/03/computational-humanities-vs-digital.html>.

MAYER-SCHÖNBERGER, Viktor et CUKIER, Kenneth, 2013, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Boston, Eamon Dolan/Mariner Books.

MONTUSCHI, Eleonora, 2003, *The Objects of Social Science*, Londres, Continuum Press.

MONTUSCHI, Elenora, 2004, « Rethinking objectivity in social science », *Social Epistemology: A Journal of Knowledge, Culture and Policy*, vol. 18, nos 2-3, p. 109-122.

MONTUSCHI, Eleonora, 2008, « Should we still compare the social sciences to the natural sciences? », *Sociologica*, nº 3. Consultable en ligne : <http://www.sociologica.mulino.it/journal/article/index/Article/Journal:ARTICLE:274/Item/Journal:ARTICLE:274>.

PLANTIN, Jean-Christophe et MONNOYER-SMITH, Laurence, 2014, « Ouvrir la boîte à outils de la recherche numérique », *tic&société*, vol. 7, nº 2, mai. Consultable en ligne : <https://ticetsociete.revues.org/1527>.
DOI : 10.4000/ticetsociete.1527

PLANTIN, Jean-Christophe, LAGOZE, Carl, EDWARDS, Paul N. et SANDVIG, Christian, à paraître, « Big data is not about the size: When data tranform scholarship », dans Clément Mabi, Jean-Christophe Plantin et Laurence Monnoyer-Smith (éd.), *Les données à l'heure du numérique. Ouvrir, partager, expérimenter,* Éditions de la Maison des Sciences de l'Homme, Paris. Consultable en ligne : <http://pne.people.si.umich.edu/PDF/Plantin%20et%20al.%202015%20Big%20Data%20is%20not%20about%20Size%20-%20pre-press%20version.pdf>.

QUETELET, Adolphe, 1835, *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*, Paris, Bachelier, imprimeur-libraire.

RAVENSTEIN, Ernst George, 1885, « The laws of migration » *Journal of the Statistical Society of London*, vol. 48, nº 2, juin, p. 167-235.
DOI : 10.2307/2979181

RAVENSTEIN, Ernst George, 1889, « The laws of migration », *Journal of the Royal Statistical Society*, vol. 52, nº 2, 1889, p. 241-305.
DOI : 10.2307/2979181

RUSSO, Federica, 2009, *Causality and Causal Modelling in the Social Sciences. Measuring Variations,* Dordrecht, Springer.

Socio, 2015, dossier « Le tournant numérique » coordonné par Dana Diminescu et Michel Wieviorka, nº 4.

STRASSER, Bruno J., 2012, « Data-driven sciences: From wonder cabinets to electronic databases », *Studies in History and Philosophy of Biological and Biomedical Sciences,* vol. 43, nº 1, p. 85-87.
DOI : 10.1016/j.shpsc.2011.10.009

WATTS, Duncan J. et STROGATZ, Steven H., 1998, « Collective dynamics of "small-world" networks », *Nature,* vol. 393, nº 6684, 4 juin, p. 440-442.
DOI : 10.1038/30918

WRIGHT, Alex, 2007, *Glut: Mastering Information Through The Ages*, Ithaca et Londres, Cornell University Press.

ZIMMER, Michael, 2010, « "But the data is already public": On the ethics of research in Facebook », *Ethics and Information Technology,* vol. 12, nº 4, p. 313-325.
DOI : 10.1007/s10676-010-9227-5

Notes

1 On citera par exemple le modèle de traitement distribué de données MapReduce.

2 Voir : <http://www.sdss.org/>.

3 Voir : <http://www.lirmm.fr/mastodons/>.

4 En effet, l'utilisation de la théorie des probabilités et de la statistique remonte à un temps plus ancien, dès les premiers recensements au XVIIe siècle, le but étant de pouvoir mesurer et quantifier différents aspects de la population (à ce sujet, voir Courgeau, 2012).

5 Le développement des modèles probabilistes a minimisé la question du déterminisme en sciences sociales ; toutefois, il faut remarquer que le problème peut se poser à nouveau lorsqu'on discute des questions concernant, par exemple, la liberté individuelle. Cela relève davantage de l'éthique ou de la philosophie politique, et se pose de manière orthogonale par rapport à la question épistémologique que nous abordons dans cet article.

6 Voir à ce sujet Russo (2009) et Illari et Russo (2014).

7 Nous laissons de côté dans cet article la dimension éthique de ces recherches, par exemple en termes de respect de la vie privée (Zimmer, 2010), de surveillance (Crawford, 2014) ou sur l'inégalité de l'accès à ces moyens de recherche (boyd et Crawford, 2012).

8 La première est la possibilité d'analyser de grandes quantités de données sur un sujet au lieu de se contenter de données restreintes. La deuxième est la volonté de saisir la complexité des données du monde au lieu de privilégier l'exactitude. La troisième est une considération grandissante envers les corrélations, au lieu d'une quête perpétuelle pour une causalité insaisissable. (Notre traduction.)

9 Pour une discussion approfondie, voir Illari et Russo (2014).

10 Plusieurs numéros thématiques parus depuis 2013 portent sur les *big data*, entre autres le *Journal of Broadcasting and Electronic Media* (2013) ; *First Monday* (2013) ; *International Journal of Communication* (2014).

11 En philosophie des sciences, Bogen et Woodward (1988) proposent une révision des termes données (*data*), phénomène, et observation afin de contrer une vue positiviste, trop simpliste par rapport à la définition et à la distinction entre observable et non observable.

12 Les perceptions d'une « surcharge d'information » (ou « déluge de données ») se sont répétées de la Renaissance jusqu'à la période moderne, et se sont accompagnées à chaque fois de l'invention de technologies pour prendre en charge cette surcharge perçue. (notre traduction.)

13 « *Reverse Engineering and other Limitations. You will not or attempt to (and will not allow others to) [...] 3) sell, rent, lease, sublicense, distribute, redistribute, syndicate, create derivative works of, assign or otherwise transfer or provide access to, in whole or in part, the Licensed Material to any third party except as expressly permitted herein* », *Developer*

Agreement & Policy, Twitter Developer Agreement, 18 mai 2015. Consultable en ligne : <https://dev.twitter.com/overview/terms/agreement-and-policy>.

14 Venture Beat : <http://venturebeat.com/2014/06/07/exclusive-to-sell-ads-in-the-developing-world-facebook-is-hiring-sociologists/>.

15 Au cours de cette étude, les chercheurs ont modifié le fil de nouvelles (*news feed*) d'une très grande quantité de comptes Facebook en affichant des éléments tantôt négatifs, tantôt positifs, afin d'en analyser les conséquences sur les comportements en ligne.

16 Traditionnellement garanties par l'approbation au préalable de l'étude par une instance de régulation publique, tels les Institutional Review Board aux États-Unis, et par la signature d'un formulaire de consentement à chaque sujet d'étude.

Pour citer cet article

Référence papier
Jean-Christophe Plantin et Federica Russo, « D'abord les données, ensuite la méthode ? », *Socio*, 6 | 2016, 97-115.

Référence électronique
Jean-Christophe Plantin et Federica Russo, « D'abord les données, ensuite la méthode ? », *Socio* [En ligne], 6 | 2016, mis en ligne le 11 mai 2016, consulté le 28 novembre 2019. URL : http://journals.openedition.org/socio/2328 ; DOI : 10.4000/socio.2328

Cet article est cité par

- Mericskay, Boris. (2019) Potentiels et limites des traces (géo)numériques dans l'analyse des mobilités : l'exemple des données de la plateforme de covoiturage BlaBlaCar. *Cybergeo*. DOI: 10.4000/cybergeo.31990

Auteurs

Jean-Christophe Plantin
Jean-Christophe Plantin est *assistant professor* en media and communications studies à la London School of Economics and Political Science. Ses travaux portent sur le projet politique de la « science des données », les méthodes numériques en sciences sociales, et les usages citoyens des applications cartographiques.

j.plantin1@lse.ac.uk

Federica Russo
Federica Russo est *assistant professor* en philosophie des sciences à l'université d'Amsterdam. Ses intérêts de recherche concernent principalement la modélisation causale en sciences sociales, biomédicale et dans le domaine du « *policy making* », aussi bien que les relations entre science et technologie.

f.russo@uva.nl

Droits d'auteur