L'ABONDANCE ET SES REVERS. BIG DATA, OPEN DATA ET RECHERCHES SUR LES QUESTIONS SOCIALES

Étienne Ollion

Caisse nationale d'allocations familiales (CNAF) | « Informations sociales »

2015/5 n° 191 | pages 70 à 79

ISSN 0046-9459

Article disponible en ligne à l'adresse :

http://www.cairn.info/revue-informations-sociales-2015-5-page-70.htm

!Pour citer cet article :

Étienne Ollion, « L'abondance et ses revers. Big data, open data et recherches sur les questions sociales », *Informations sociales* 2015/5 (n° 191), p. 70-79.

L'abondance et ses revers. Big data, open data et recherches sur les questions sociales

Étienne Ollion – sociologue

Objet de nombreuses attentions, les big data sont aussi l'objet de controverses polarisées. Plutôt que de les trancher dans ces débats, cet article propose un tour d'horizon des principaux arguments évoqués. Force du nombre, représentativité, protection des enquêtés ou concurrence nouvelle dans l'énonciation d'un discours sur le monde social sont certains des enjeux que rencontrent aujourd'hui les chercheurs.

Il est devenu difficile d'échapper aux big data. Chaque jour ou presque, ces vastes ensembles de données sont au cœur de l'actualité. Issues des objets connectés, des capteurs que nous que nous utilisons de manière croissante, et plus généralement de la numérisation généralisée du quotidien, ces informations massives sont régulièrement louées. Les entreprises sont toujours plus nombreuses à se réorganiser pour tenter de valoriser ces données. Dans la lignée du mouvement pour l'ouverture des données (open data), les administrations sont invitées à s'en saisir, que ce soit pour rationaliser leur pratique ou pour améliorer leur service au public. Et partout, les discours prophétiques se multiplient, qui mettent l'accent **sur le potentiel scientifique, économique et social de ces données.**

Cette abondance nouvelle n'est cependant pas allée sans critiques. Les dangers de cette numérisation du quotidien ont ainsi été rapidement pointés. Les différents scandales, qui ont vu des États espionner massivement les populations, avec ou sans autorisation légale, le rappellent régulièrement. L'accumulation d'informations et le croisement des différentes sources permet de connaître

des pans entiers de la vie des individus, parfois les plus personnels. La notion de vie privée est ainsi remise en cause. L'accumulation de données sur les pratiques de consommation des personnes permet de connaître les pratiques avec une granularité inconnue jusqu'alors. Ces informations sont utilisées par de **nombreuses entreprises à des fins commerciales – quand elles ne circulent pas** sur le web suite à une faille de sécurité.

Dans le domaine de la recherche, les mêmes oppositions se rejouent, réfractées. **Plusieurs chercheurs annoncent qu'une révolution scientifique est en cours. La** masse de données, leur granularité voire simplement leur existence doit permettre de poser à nouveau des questions non résolues d'hier, mais aussi d'en aborder de nouvelles. Certains annoncent même la découverte prochaine de « lois du social » sur la base de ces méthodes (Pentland, 2014). D'autres, moins enthousiastes, rappellent que la multiplication des données ne donne pas forcément lieu à un surcroît de savoirs (Abbott, 2014), et que ceux qui prophétisent des changements de **paradigme suite à cet afflux oublient peut-être un peu vite les leçons de l'histoire.**

Plutôt que de trancher la question trop générale de la désirabilité des big data ou **de l'open data pour la recherche (des termes eux-mêmes trop flous pour permettre** une discussion précise), cet article propose un tour d'horizon des principaux arguments avancés. Sur cette base, il montre ensuite que les changements ne sont peut-être pas là où on les attend, et qu'en fait de révolution toujours à **venir, des décalages se produisent, qui reconfigurent la manière dont se fait la** recherche dès maintenant.

Des promesses en nombres

Du point de vue des sciences sociales, la multiplication des données numériques est souvent décrite comme riche d'une triple promesse (Ollion et Boelaert, 2015). Premièrement, du point de vue *empirique,* l'augmentation du nombre de données devait permettre une meilleure connaissance de certains sujets, ainsi **que l'exploration d'autres qui étaient jusque-là difficiles à analyser. Pour des** sociologues intéressés aux questions de mobilité, l'existence de capteurs sur les **cartes des usagers des transports en commun offre des informations d'une finesse** comme d'une extension inimaginables jusqu'alors. Centralisées dans d'immenses bases de données, elles permettent de dresser une carte autrement plus précise des déplacements quotidiens d'une population. Ces informations sont d'autant plus intéressantes que, contrairement aux enquêtes classiques, elles ne se périment pas. Ou, plutôt, la facilité de la captation permet leur mise à jour permanente.

Dans un second temps, du point de vue *méthodologique,* les avancées potentielles **sont tout aussi significatives. En raison de l'automatisation de la collecte de** certaines informations, peu de personnes échappent, sur de nombreux sujets, à la

connaissance de l'enquêteur. La différence est sensible avec la pratique classique de la recherche. Faute de pouvoir interroger l'ensemble des enquêtés ou coder tous les dossiers dont ils disposaient, les chercheurs devaient souvent faire des choix et en sélectionner quelques-uns. Cette procédure d'échantillonnage ne se pose plus avec le recours aux méthodes numériques. La collecte et le traitement des données pour l'ensemble de la population ne prennent pas plus de temps que pour une simple portion de celle-ci. En Belgique, les sociologues de l'université peuvent ainsi, depuis 1999, accéder à l'ensemble des informations produites par les services de protection sociale du pays. Cette collaboration, initiée à des **ins d'évaluation des politiques publiques, permet aussi l'accès à des données** quasi exhaustives sur l'ensemble de la population (Knapen *et al.*, 2014). Parce qu'elles croisent des informations de sources diverses, ces informations brossent aussi un tableau riche des modes de vie et de leurs transformations. Et ici comme ailleurs, la complétude des données fait que la très classique question de la représentativité cesse de se poser, puisque l'échantillon tiré est de la même taille que la population (« N = tous »).

> Une « physique sociale »

Troisième promesse : cette abondance est souvent décrite comme le point de départ de renouveaux théoriques. Les auteurs d'un ouvrage grand public sur le sujet écrivent ainsi que *« les effets [des big data] vont modifier la manière dont nous nous percevons. (…) Les big data vont modifier les humanités en profondeur, transformer les sciences sociales »* **(Aiden et Michel, 2013, p. 8), une af i rmation** également avancée par d'autres (Cukier et Mayer-Schönberger , 2014). Cette ambition généralisatrice et les espoirs placés dans ces données sont bien exprimés par l'un des pionniers de ces techniques, Alex Pentland, qui a publié récemment un ouvrage intitulé *Physique sociale* (2014), dans lequel il prophétisait un changement de paradigme dans les sciences sociales. La raison de ce changement se trouve dans la masse de données, qui doit permettre de faire émerger des régularités, mais aussi trouver des lois du fonctionnement du monde social. Invisibles des acteurs, mais aussi des chercheurs en sciences sociales trop proches de leurs objets pour discerner les tendances de fond qui régissent les sociétés, elles seraient accessibles aux informaticiens qui manipulent de vastes ensembles et leur imposent des traitements statistiques avancés.

> *“(…) cette abondance est souvent décrite comme le point de départ de renouveaux théoriques.”*

Les infortunes du volume

Las, le moins que l'on puisse dire est que les promesses n'ont pas toutes été tenues, et beaucoup ont même été déçues dans le domaine des recherches sociales. Plusieurs raisons, qui suivent la tripartition évoquée précédemment, expliquent que les résultats soient moins parlants que prévus. La première, empirique, est que l'augmentation du nombre des données disponibles ne doit

pas être confondue avec leur utilité pour la recherche[1]. Produites pour les besoins du fonctionnement d'un service, qu'il soit administratif ou commercial, les informations désormais disponibles en masse se révèlent souvent éloignées des questions que se posent les chercheurs. Il arrive aussi régulièrement que les **données, pour massives qu'elles soient, se révèlent i nalement très frustes par** rapport aux connaissances déjà établies. L'exemple évoqué ci-dessus l'illustre bien : les données de mobilité des usagers des transports en commun. Chaque trajet donne désormais lieu, pour les porteurs d'une carte d'abonnement, à un enregistrement de la date et du transport choisi. En cela, elles pourraient être considérées comme fécondes pour une recherche sur les mobilités urbaines. Ces données « libérées » se révèlent toutefois assez pauvres pour qui souhaite en savoir plus sur les usagers, dont les informations personnelles sont – heureusement – indisponibles. Faute d'apprendre quoi que ce soit sur les utilisateurs ou de pouvoir les suivre (changent-ils de ligne ou de moyen de transport ?), les usages restent limités. On doit souvent se contenter de visualisations sur les trajets d'un point à un autre du réseau (plutôt que sur les trajets réels des utilisateurs), ou d'informations sur les endroits les plus fréquentés – ce qui est d'ailleurs l'objectif, puisqu'elles sont mises à disposition avant tout en vue de la création de services dédiés *via* des applications.

> Big data ne veut pas toujours dire *rich data*

Big data ne signi i e pas donc automatiquement « rich data ». Ces données ne sont pas forcément de bonne qualité non plus. Parmi les nombreuses informations qu'on trouve sur les portails d'open data mis en place par les entreprises ou les administrations, on trouve parfois des données à la provenance comme à la qualité douteuses. Si les données de capteurs, les traces laissées par les utilisateurs lors de leurs passages sur Internet nous en apprennent parfois peu, l'automaticité de l'enregistrement assure une certaine harmonie entre elles. Ce n'est toutefois pas toujours le cas. Les données produites par des administrations enjointes par les pouvoirs publics à « l'ouverture » font que les bases de différents services **sont rapidement agrégées a i n de répondre à cette commande politique. Cela** se fait souvent sans les habituelles procédures de contrôle et d'harmonisation que mettent en place les chercheurs dans leurs travaux, et souvent sans même une information sur les conditions de la production des données. Mais une fois mises en ligne sur des sites spécialisés et stockées sous forme de tableurs, ces informations composites semblent tout aussi objectives que celles produites **patiemment et avec des véri i cations. Elles sont d'ailleurs utilisées comme telles.**

Du point de vue méthodologique, la quantité ne saurait être un argument non plus. Car si on peut potentiellement disposer de l'ensemble des informations, ce **cas de i gure est plutôt l'exception que la règle. Dans le reste des cas, on dispose** de plus de données, mais pas de toutes. Il ne faut alors pas oublier la leçon de méthodologie statistique, selon laquelle un échantillon bien construit est souvent plus probant qu'un amas d'informations dont on ne connaît pas la relation à

la population d'ensemble. Cette supériorité, bien connue des statisticiens, fut démontrée par George Gallup à l'occasion de l'élection présidentielle de 1936 aux États-Unis. Gallup, qui venait de fonder une société de sondage, avait prédit la victoire du président sortant Franklin D. Roosevelt contre l'ensemble des commentateurs politiques mais aussi, et surtout, contre la célèbre revue intellectuelle *The Literary Digest*. Cette dernière avait en effet annoncé la victoire du candidat républicain Landon sur la base des déclarations faites par ses lecteurs, dont plusieurs centaines de milliers avaient renvoyé un coupon à remplir inséré dans la revue. Gallup avait, quant à lui, constitué un échantillon de quelques milliers d'électeurs seulement (Didier, 2009) qui non seulement ne comportait pas les biais dont souffrait la base du *Digest* (sélection d'une petite partie de l'électorat, auto-sélection *via* le choix de la réponse), mais était constitué de manière à être représentatif de la population en âge de voter. La leçon de théorie des sondages vaut toujours à l'heure actuelle : on approche souvent mieux les choix d'une population avec un échantillon bien structuré de 1 000 personnes qu'avec un tirage sans principe de plusieurs centaines de fois ce chiffre.

Enfin, les espoirs théoriques ont été, eux aussi, déçus jusqu'alors. Les déclarations qui annonçaient une nouvelle physique sociale sont, à l'heure actuelle en tout cas, restées lettre morte. Fantasme récurrent, régulièrement prophétisée depuis le milieu du XIX[e] siècle et la naissance des sciences humaines et sociales, la découverte de grandes lois du social ne s'est pas vraiment concrétisée. La situation actuelle n'échappe pas à la règle. Paradoxalement, la multiplication des données numériques semble même avoir déplacé la recherche vers un côté plus descriptif. L'abondance de données a pour l'instant donné lieu à des publications qui accumulent les informations empiriques plutôt qu'elles ne donnent à voir de grandes régularités du social.

Décalages

Biais, qualité douteuse, informations abondantes mais pas toujours utiles... L'avalanche de données numériques ne produira certainement pas le changement radical prophétisé par certains. Le caractère imparfait des données comme **l'absence de révolution scientifique ne doivent toutefois pas nous empêcher de** saisir les opportunités offertes par la multiplication des données numériques, qu'elles soient « big » ou pas, numériques ou pas. Elles ne doivent pas non plus empêcher de saisir les changements à l'œuvre dans la pratique de la recherche. Parmi ceux-ci, deux sont particulièrement saillants, qui travaillent les sciences sociales dès à présent.

> Réorganisations disciplinaires

Le premier changement a trait aux compétences requises. Pour abondantes qu'elles soient, les informations ne sont jamais immédiatement disponibles. Plutôt que « données », elles sont plutôt acquises, à travers un processus de collecte et de nettoyage qui peut être long. Pour ce faire, les chercheurs doivent

connaître l'informatique, au moins *a minima*. Il ne faut pas croire que celle-ci ne sert que ceux qui travaillent sur Internet et étudient les interactions en ligne. Par exemple, une historienne qui travaillerait sur les débats à l'Assemblée nationale en utilisant des fichiers scannés peut vouloir les transformer en fichiers texte, afin de sélectionner l'information pertinente, faire des recherches, voire opérer des traitements (quantitatifs ou non). Même quand ils n'ont pas à produire eux-mêmes leur base de données, les chercheurs ont le plus souvent besoin de la préparer avant de pouvoir l'exploiter. Une économiste qui travaille sur des données fiscales peut avoir à sélectionner des informations dans l'ensemble fourni par l'administration avant de pouvoir les traiter statistiquement (supprimer les scories, repérer les doublons, les erreurs de codage, harmoniser…). Une sociologue peut vouloir modifier le format d'un fichier afin de l'utiliser dans un autre logiciel…

L'importance de cette compétence informatique, que l'on voit d'ores et déjà poindre dans les fiches de postes comme dans les formations en sciences sociales, a de nombreux effets. Elle favorise les collaborations interdisciplinaires entre les chercheurs en sciences sociales et ceux en sciences de l'informatique. De fait, sur de nombreux objets, une véritable interdisciplinarité est en train de se mettre en actes. Stimulante, elle ne va pas sans poser de question aux chercheurs et à leurs habitudes. L'un des déplacements les plus saillants tient certainement à l'accroissement de la division du travail scientifique : ceux qui codent les données ne sont plus ceux qui extraient les données, lesquels ne sont parfois pas ceux qui les traitent. L'allongement de ces chaînes d'interdépendances modifie le fonctionnement des disciplines, tout en mettant l'analyste à une distance plus grande du terrain.

Par ailleurs, en raison de la disponibilité de données, les chercheurs en sciences sociales sont désormais concurrencés par d'autres groupes professionnels. Jusque-là en situation de quasi-monopole pour l'énonciation de vérités quantitatives sur le monde social, les économistes, les sociologues ou les démographes voient leur juridiction concurrencée – par exemple par des data-journalistes, ces spécialistes du recueil et de la visualisation des données, dont la présence croissante dans les rédactions limite le recours aux chercheurs en sciences sociales, en tout cas en ce qui concerne la production de nombres. Certaines des demandes adressées hier aux chercheurs sont désormais traitées en interne, ce qui n'est pas sans entraîner des modifications sur la manière dont les sujets sont abordés, que ce soit en termes d'objets, d'approches ou de regards portés.

> Accès et gestion des données

Une autre évolution qui concerne les chercheurs tient à l'accès et à la gestion des données. L'abondance des informations existantes ne signifie pas en effet que ces dernières sont librement accessibles, ni même facilement utilisables. L'inverse

preuve, et le fait que l'activité scienti i que ne se limite pas à tester des hypothèses bien formulées mais progresse par essais, échecs et reformulations. Ne collecter **que des données dé i nies comme pertinentes** *ex ante* ou imposer un protocole strict en amont de l'enquête, tout cela revient à n'utiliser les données que pour valider des hypothèses déjà présentes. Dans bien des cas, c'est, tout simplement, s'interdire de découvrir.

Quand on sait le lobbying actif que réalisent plusieurs entreprises privées pour avoir le droit de traiter les données individuelles, non seulement pour faire de la publicité en ligne ou offrir des services sur mesure (ce qui, dans **le cas d'assurances santé, peut donner lieu à une tari i cation en fonction de** l'histoire médicale du patient), et quand on sait que ces données sont perdues, échangées voire revendues entre prestataires, ces mesures de précaution sont compréhensibles. Elles sont toutefois tellement strictes qu'elles sont contournées en permanence. L'excès de rigueur donne donc lieu à une absence de régulation effective, alors même que la granularité des informations collectées tout comme la multiplication des données (celles-ci étant souvent stockées sur des supports mobiles – portables, disques durs externes – ou dans le *cloud*) nécessitent une réorganisation complète de la manière dont travaillent les chercheurs. Droit d'enquêter d'un côté, protection des enquêtés de l'autre, voilà les deux aspects, lesquels ne s'opposent pas forcément, que les chercheurs doivent concilier de manière pressante (Laurens et Neyrat, 2010). De ce point de vue, la mise en place d'une « exception recherche » doublée d'une organisation interne semble être une voie aussi prometteuse que nécessaire.

C'est que le chercheur se trouve face à un autre enjeu pressant, et ce d'autant plus qu'il travaille sur des questions sociales : celui de la démarcation entre son travail et celui du policier, du magistrat ou du travailleur social. Les importants volumes de données sont en effet exploités de manière croissante par les administrations pour détecter les fraudes. Plus généralement, la collecte et l'analyse de volumes de données désormais disponibles servent des organisations, publiques ou privées, dans leur recherche d'informations sur des individus (Harcourt, 2014). Le fait qu'elles soient de qualité douteuse, périmées et parfois inutilisables est **i nalement moins important que le sentiment de contrôle généralisé ressenti de** manière croissante par les populations. Et les chercheurs, qu'ils prennent part directement à des contrôles ou pas, sont rapidement associés à des enquêteurs comme les autres, ce qui n'est pas sans conséquences pour les enquêtes de terrain qu'ils mènent.

Des volumes inouïs ?

« La quête humaine du savoir et le travail académique sont entravés par de nombreux obstacles, au premier rang desquels se trouve l'abondance de travaux disponibles ». Cette remarque, faite au XIV^{e} siècle par le philosophe arabe Ibn

Khaldun dans la *Muqaddimah* (1969, p. 414), rappelle utilement que la phase actuelle présente des similitudes frappantes avec d'autres époques. Plusieurs fois au cours des siècles passés, les sociétés ont ainsi expérimenté le sentiment d'abondance dans le domaine des connaissances. Ce fut ainsi le cas lors de l'invention de l'imprimerie et de la diffusion des imprimés, ou au moment de l'apparition des premiers ordinateurs dans les années 1950 (une époque où certains annonçaient l'avènement d'une « *big science* » proche de la situation actuelle). À chaque fois, des débats ont eu lieu entre prophètes d'une révolution à venir et critiques de la surcharge informationnelle et de ses dangers pour la **connaissance. De fait, chaque changement technologique signi i catif donne** lieu à des fantasmes autant qu'à des mutations effectives. L'abondance nouvelle de données numériques est l'un de ces changements, et ses effets sont déjà sensibles : au-delà des débats, les disciplines et leurs frontières changent, les **objets évoluent, tout comme le rapport au terrain peut être modi i é. Si des** déplacements ont bien lieu, les conséquences de long terme du mouvement actuel sont, elles, imprévisibles. Elles dépendent largement de l'orientation que nous lui donnerons.

Notes

1 – Voir Abbott, 2014, pour une critique cinglante de cette illusion.

2 – En France, l'article 112-3 du Code de la propriété intellectuelle punit de 500 000 euros d'amende et de trois ans d'emprisonnement ces actes.

3 – http://ameli-direct.ameli.fr

Bibliographie

■ Abbott A., 2014, *Digital Paper. A Manual for research and writing with library and internet research,* University of Chicago Press.

■ Aiden E. et Michel J.-B., 2013, *Uncharted: Big Data as a Lens on Human Culture,* Riverhead Books.

■ Cukier K. et Mayer-Schönberger V., 2014, *Big Data: A Revolution That Will Transform How We Live, Work, and Think,* Eamon Dolan/Mariner Books.

■ Didier E., 2009, *En quoi consiste l'Amérique ? Les statistiques, le new deal et la démocratie,* La Découverte.

■ Harcourt B., 2014, Governing, Exchanging, Securing: Big Data and the Production of Digital Knowledge, *Columbia Public Law Research Paper n° 14-390,* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2443515 (consultation le 1er mai 2015).

■ Ibn Khaldun, 1969, *The Muqqadimah. An Introduction to History,* Princeton University Press, [1377].

■ Knapen H., Braes S., Ermans T. et Erremans W. (dir.), 2014, *Het Datawarehouse, een duizendpoot! Perspectieven van het Datawarehouse Arbeidsmarkt en Sociale Bescherming,* Gand, Academia Press.

■ Laurens S. et Neyrat F, 2010, *Enquêter : de quel droit ? Menaces sur l'enquête en sciences sociales,* Vulaines-sur-Seine, Éditions du Croquant.

■ Ollion É. et Boelaert J., 2015, Au-delà des big data. Les sciences sociales face à la multiplication des données numériques, *Sociologie,* vol. 6, n° 3.

■ Pentland A., 2014, *Social Physics. How good ideas spread – The lessons from a new science,* New York, Penguin.