DIGITAL HUMANITIES

Direction de la prospective et du dialogue public

BIG DATA ET SCIENCES SOCIALES

RÉVOLUTION, ILLUSION OU EXTENSION ?

Rapport de Boris CHABANEL - NOVA 7

Février 2016

NOVA 7

GRANDLYON la métropole

SOMMAIRE

1. De quoi parle-t-on ?

Du « big data »…

La notion de big data fait référence à la conjonction de deux phénomènes qui résultent des progrès[1] et de la diffusion des technologies de stockage et de calcul issues de la révolution numérique (Mayer-Schönberger, 2014 ; Boullier, 2015 ; Ollion et Boelaert, 2015 ; Ganascia, 2015) :

⇒ une **mutation quantitative et qualitative des données** produites par les entreprises, les particuliers, les acteurs publics, les scientifiques :

- Collectées sous forme numérique : au cours des 15 dernières années, nous sommes passés d'un monde dans lequel les trois quarts des données étaient analogiques à un monde où celles-ci représentent à peine 1 % du total.
- Volume : les volumes des données collectées auraient été multipliés par 100 entre 1987 et 2007, et doublent depuis tous les ans ou tous les deux ans.
- Vélocité : les données tendent à être collectées et analysées en continu et en temps réel, et dès lors que l'analyse de ces données s'effectue elle-même en temps réel il devient possible de ne plus stocker les données mais de les analyser en flux.
- Variété : les données analysées ne sont plus forcément structurées comme dans les bases de données classiques, mais peuvent être du texte, des tableaux, des images, des sons, des traces numériques…

⇒ le **développement d'outils informatiques** capables de stocker et de traiter de gigantesques volumes de données. Il s'agit en particulier du « machine learning », forme d'intelligence artificielle qui croise mathématique et informatique afin de réaliser des tâches qui ne pouvaient pas être effectuées jusqu'à présent par les méthodes statistiques traditionnelles du fait de la complexité et de la masse des données disponibles. Le machine learning (ou apprentissage automatique) est une méthode d'analyse de données qui mobilise des algorithmes[2] capables d'apprendre de façon itérative à partir de données et sur la base d'un corpus d'exemples pertinents, sans être programmée explicitement où chercher.

… aux « computational social sciences »

Selon Wikipedia, les sciences sociales computationnelles (Computational social science) désignent des sous-disciplines académiques concernées par l'application des approches informatiques aux sciences sociales. Cela signifie que les outils

---

[1] *S'agissant de l'accroissement des performances des technologies numériques de l'information et de la communication, les analyses font souvent mention de plusieurs « lois » (Delort, 2015). La plus connue, la loi de Moore, qui prévoit un doublement de la densité d'inscription sur puce de silicium tous les 18 mois, permet un accroissement incessant de la capacité de calcul. La loi de Kryder concerne le stockage sur disque magnétique avec un doublement de densité tous les 13 mois. La loi de Nielsen, celle de la fonction transmettre, fait doubler tous les 21 mois la capacité des réseaux publics.*

[2] *Un algorithme permet de résoudre un problème en le décomposant en une suite d'instructions ou d'opérations. Il s'apparente à une procédure, mais ses composants doivent être précisément définis. C'est à cette condition qu'il peut être exécuté par des machines informatiques grâce à son écriture sous forme de code (Boullier, 2015)*

informatiques sont utilisés pour modéliser, simuler et analyser les phénomènes sociaux. On distingue les sciences économiques computationnelles (computational economics) et la sociologie computationnelle (computational sociology).

La sociologie computationnelle est une branche de la sociologie qui utilise des simulations informatiques, l'intelligence artificielle, les méthodes statistiques complexes, et les approches analytiques tels que l'analyse de réseau social, afin de développer et tester des théories de processus sociaux complexes grâce à la modélisation des interactions sociales. Elle vise la compréhension des agents sociaux, l'interaction entre ces agents, et l'effet de ces interactions sur l'ensemble social.

2. Révolution : mettre au jour les lois cachées du monde social et prédire l'avenir

Un premier ensemble de réflexions convergent autour de l'idée que l'émergence du big data constituerait une nouvelle révolution scientifique, à l'image du passage au siècle dernier de la notion newtonienne d'absolu à la relativité d'Einstein. Paru en 2014 et intitulé « social physics », le dernier ouvrage d'Alex Pentland, un des fondateurs et principaux chefs de file du courant des « computational social science » et chercheur au Media Lab du MIT, défend l'idée que le big data va rendre possible la modélisation mathématique de la société, c'est-à-dire la mise au jour des lois régissant son fonctionnement (H.Guillaud, 2014). Cette révolution découlerait de la profusion des données mis à disposition par les technologies numériques et de la mise en œuvre d'une méthode d'analyse radicalement inductive. A terme, cette perspective rendrait obsolète l'approche hypothético-déductive des sciences sociales.

Une profusion de données numériques qui offre une nouvelle vision du monde

Jusque récemment, la collecte de données était difficile, chronophage et coûteuse. Les données récoltées restaient incomplètes, déclaratives… Les scientifiques se sont adaptés à cette rareté de la donnée en développant des méthodes de recueil et d'analyse pouvant fonctionner avec peu de données. Par exemple, la méthode à peine centenaire de l'échantillonnage aléatoire utilise ingénieusement une petite partie pour avoir une idée du tout (Mayer-Schönberger, 2014).

Aujourd'hui, l'informatisation croissante des organisations, le développement d'internet et de la téléphonie mobile, la multiplication des capteurs de toute sorte, ou encore la numérisation croissante du quotidien paraissent en mesure de remédier à cette lacune en permettant de collecter et d'accéder à des flux de données inédites, précises et en temps réel sur des millions de personnes. De fait, nous avons aujourd'hui à notre disposition beaucoup plus de données que jamais dans l'histoire humaine.

La possibilité d'accéder à d'immenses ensembles de données a été accueillie avec enthousiasme par beaucoup de chercheurs, pour plusieurs raisons (Ollion et Boelaert, 2015).

⇒ Du point de vue empirique, le big data laisse augurer la possibilité d'étudier des sujets jamais appréhendés jusque-là, et surtout de connaître avec une précision infiniment plus grande des aspects entiers du comportement humain.

L'exemple des études de mobilité

Les études de mobilité bénéficient de la collecte régulière et automatique de données via des capteurs RFID, ces puces de radio-identification qui enregistrent les informations relatives aux objets – ou aux personnes, aux animaux – auxquelles elles sont attachées et peuvent les transmettre régulièrement vers un serveur qui les collecte. Installées sur les

cartes d'abonnement de transports en commun, elles fournissent un tableau jusque-là inaccessible des mobilités urbaines. Les données de téléphone portable sont encore plus précises. Il devient ainsi possible de connaître les mobilités de populations entières avec un luxe de détails impensable jusqu'alors.

⇒ Sur un plan méthodologique, les données numériques auraient pour avantages d'être plus faciles à stocker, à retrouver et à analyser, et leur actualisation (collecte en continu et en temps réel) permettrait de limiter le risque classique de la péremption que connaissent les enquêtes par questionnaire. De même, l'automatisation des enregistrements et l'accélération des traitements permettraient de s'affranchir des contraintes classiques d'échantillonnage de la population que l'on cherche à analyser : sur de nombreux sujets, il est désormais quasiment aussi rapide d'étudier la population entière qu'un échantillon restreint. Le changement est significatif, puisque c'est la question de la représentativité qui cesse de se poser quand l'un et l'autre sont similaires : les statisticiens disent « N = All ».

Une révolution épistémologique : la remise en cause des catégories d'analyse a priori

Cette formidable augmentation quantitative des données disponibles entrainerait un changement qualitatif majeur sur le plan théorique. Le Big Data permettrait en effet de ne plus plier la réalité à des catégories définies *a priori* en laissant les données fournir elles-mêmes les catégories qu'elles contiennent (Mayer-Schönberger, 2014). Autrement dit, l'abondance nouvelle de données permettrait la découverte de régularités inconnues jusqu'alors, la mise au jour des lois de fonctionnement du monde social (Ollion et Boelaert, 2015).

Traditionnellement, les sciences expérimentales partent d'une théorie qu'elles confrontent à la réalité. Cette confrontation passe par la construction de dispositifs matériels d'observation ; selon qu'il y a écart ou non entre ces observations et ce que la théorie anticipe, celle-ci se trouve invalidée ou validée. Les big data révoquent ce schéma hypothético-déductif classique au sens où la théorie n'a plus la primauté (Ganascia, 2015). Les dispositifs de recueil d'observations ne sont pas conçus au regard d'une hypothèse que l'on cherche à prouver. La collation des données se fait automatiquement, sans idée a priori. On exploite ensuite ces données à l'aide d'outils d'intelligence artificielle et d'apprentissage machine en procédant à des expériences permettant de tester systématiquement de grands nombres d'hypothèses sur les données. En résumé, par rapport aux statistiques classiques, les big data se caractérisent par l'absence de théorie initiale et de protocole de collecte, et visent principalement non pas à valider des causalités formulées a priori mais à trouver des corrélations apparentes entre les données.

Ainsi, selon Viktor Mayer-Schönberger, professeur de gouvernance et de régulation de l'internet à l'université d'Oxford, le big data aurait la vertu de donner à voir beaucoup plus précisément « ce qui se passe » sans nous en remettre à des approches et préconçues et simplificatrices du « pourquoi ». A l'appui de sa réflexion, il convoque les analyses du prix Nobel d'économie Daniel Kahneman qui ont mis en évidence le fait que les hommes aiment se représenter le monde en termes de causes et d'effets mais que nos intuitions causales immédiates sont souvent fausses. Dès lors, la logique de causalité n'aurait plus le monopole de la

recherche de sens, celle-ci pouvant être alimentée désormais par les analyses corrélationnelles du Big Data.

Fin de la théorie, fin des sciences sociales ?

Le journaliste scientifique Chris Anderson avait lancé un pavé dans la mare en publiant en 2008 un court article au titre pour le moins explicite : « The end of theory : the data deluge makes the scientific method obsolete ». Son argumentation se fondait sur le constat que, dans plusieurs domaines scientifiques tels que la génomique ou la physique, l'usage massif des données à des fins de découverte de corrélation avait remplacé selon lui la nécessité des modèles, des hypothèses et des épreuves construites pour les tester. Comme le souligne le sociologue Dominique Boullier (2015), Directeur du Social Media Lab à l'Ecole Polytechnique Fédérale de Lausanne (EPFL), les sciences sociales elles-aussi pourraient également voir leur autorité remise en cause dès lors que la collecte et le stockage de données de tous types semblent sans limite et que la puissance de calcul des machines permet de tester des milliers de corrélations sans restriction et d'en évaluer la robustesse statistique avant même que toute interprétation soit nécessaire. Autrement dit, cette forme d'« automatisation de l'induction » par essai/erreur à haute fréquence tend à rendre toute théorisation (autre qu'informatique et statistique) assez vaine, puisqu'elle n'a pas de pouvoir discriminant *a priori* par rapport à d'autres corrélations. Cela n'empêche pas le travail d'interprétation mais les hypothèses sur lesquels il se fonde tendent à être générées par les corrélations elles-mêmes.

Mais la remise en question des sciences sociales classiques ne s'exprime pas seulement au plan méthodologique. Elle se joue également en matière d'aide à la décision. Le big data permettrait en effet d'atteindre le graal : prédire l'avenir (Pentland, 2014). Pour s'en convaincre, plusieurs auteurs soulignent que les caractéristiques des données (volume, vélocité, variété) permettent de passer d'analyses basées sur des moyennes et des stéréotypes à des analyses fondées sur l'individualisation de la statistique, c'est-à-dire le calcul de profils et d'interactions individuels (Ibekwe-Sanjuan, 2014 ; H.Guillaud, 2014). D'autre part, les corrélations révélées par le machine learning, c'est à dire la prédiction de l'évolution d'une variable (quantitative ou qualitative) à partir d'un ensemble de variables explicatives, ouvre la voie à la prédiction des actions futures des individus, des groupes et, in fine, de la société à partir des données récoltées sur les actions passées (Ibekwe-Sanjuan, 2014).

Dominique Boullier alerte sur le fait que les analyses issues du big data, parce qu'elles permettent d'anticiper, avec un certain degré de certitude, des comportements ou des besoins, suscitent un intérêt croissant auprès des décideurs privés comme publics. Côté entreprises, il en veut pour preuve les pratiques des géants du numériques : Google, Apple, Facebook, Amazon, Twitter, etc. Ces plateformes constituent de puissants outils de gestion de marques en permettant de suivre en temps réel l'évolution de leur réputation ou de leur notoriété, ainsi que l'influence de leurs actions sur le public ; et ce sans passer par les interprétations et les modèles des sciences sociales. Or, ce qui préoccupe avant tout ces marques ne sont pas des données structurées et construites pour tester des causalités par exemple, mais bien des traces, qui fonctionnent comme indices et alertes, même approximatifs, non pas au niveau individuel mais au niveau de tendances, de trends. De même, ce n'est pas la réflexivité qui est recherchée mais avant tout la réactivité, la capacité à déterminer sur quel levier agir en fonction des dimensions de la marque qui sont affectées.

L'exemple d'Amazon (Mayer-Schönberger, 2014)

À l'origine, Amazon fondait ses recommandations de produits sur des catégories préconçues de consommateurs utilisées par ses experts en marketing, avec pour résultat l'impression de faire, selon un employé de l'entreprise, « ses achats avec l'idiot du village ». Quand Amazon a commencé à utiliser l'analyse du Big Data pour dégager les similitudes entre produits vendus, la pertinence de ses recommandations s'est accrue de façon vertigineuse. Le système de recommandations d'Amazon représenterait aujourd'hui environ un tiers des revenus de l'entreprise.

Les potentialités du big data commencent également à résonner sur le versant public. Les administrations de nombreux pays sont désormais dotées d'un responsable de haut rang dédié à cette thématique (Ollion et Boelart, 2015 ; Grosdhomme Lulin, 2015). De plus, selon Dominique Boullier, les méthodes des marques tendent également à prendre le dessus et à imposer leur rythme à la vie politique. La spirale de la réactivité et de l'addiction aux tweets suggèrent que nous serions entrés dans l'ère du « high frequency politics » à l'image du high frequency trading de la finance spéculative.

L'exemple de la police prédictive (Grosdhomme Lulin, 2015)

Les pratiques de predictive policing développées dès 2008 par la police de Los Angeles sous la direction de Bill Bratton sont désormais en voie de généralisation aux États-Unis, mais aussi en Allemagne, en Suisse et au Royaume-Uni. La police prédictive consiste à collecter de multiples données sur les circonstances et modalités des crimes et délits passés, à les modéliser, à les confronter en temps réel aux données géolocalisées permettant de caractériser la situation de telle rue, tel quartier ou telle zone d'un territoire, en sorte de prévoir le risque d'occurrence des faits que l'on cherche à combattre et à prépositionner en conséquence des patrouilles de police. Dans tous les cas, les gains d'efficacité constatés après quelques mois ou trimestres de mise en œuvre de ces méthodes s'avèrent spectaculaires : par exemple, les villes d'Atlanta et Los Angeles ont réussi à réduire de 20 % à 40 % certains crimes spécifiques en une année.

Au total, selon Dominique Boullier, si le souci de l'action à court terme devait l'emporter, le pouvoir d'attraction du big data auprès des décideurs publics et privés pourraient entrainer un relatif désintérêt pour les modèles explicatifs de grande portée développés par les sciences sociales. Dans ce scénario, nous passerions du lien savoir/pouvoir à un lien données/action.

3. Illusion : le big data ne peut pas tout

On le voit, l'émergence du big data amène à réinterroger les pratiques des sciences sociales. Amenés à se positionner face au caractère prétendument révolutionnaire du big data, un certain nombre de chercheurs se montrent plus dubitatifs. Selon eux, les promesses affichées ne doivent pas faire oublier un certain nombre de limites bien réelles (Ollion et Boelaert, 2015)

Les données restent faillibles

La première faiblesse du big data réside dans un défaut de questionnement des propriétés et de la validité des données numériques recueillies. Seule compterait leur disponibilité.

Or, les données issues du big data, comme toute forme de données, ne sont jamais brutes, jamais pures, jamais neutres. Parce qu'elles découlent en réalité d'un complexe processus de collecte, de nettoyage et de filtrage avec de nombreux biais, les données ne peuvent jamais être utilisées sans une opération réflexive destinée à en connaitre les conditions de production (Ollion et Boelaert, 2015 ; Ibekwe-Sanjuan). Autrement dit, le big data ne supprime pas la question cruciale de la fiabilité des données. A cet égard, le big data renvoie à des données d'origines diverses qu'il convient de distinguer afin de mieux évaluer les intérêts propres et les limites de chacune d'elle (Ollion et Boelaert, 2015).

⇒ données de l'internet : informations collectées en ligne (relatives à des pratiques qui se déroulent online : par exemple une requête sur Google) ou auxquelles on accède via le web (qui rendent compte de pratiques offline : par exemple la présence de commerces dans tel quartier). La distinction est importante car la notion de données numériques est parfois entendue comme ayant trait spécifiquement à des pratiques se déroulant en ligne.

⇒ données produites par des organisations (administrations, entreprises, associations) dans le cadre de leur fonctionnement. Elles recensent de très nombreuses informations : membres, commandes et clients, budgets, détails de l'activité des services, etc. De longue date, elles ont constitué une source utile pour les chercheurs en sciences sociales. Ces données sont propriétaires pour la plupart, de plus en plus crées en ligne et sont stockées dans les serveurs des entreprises qui les gèrent.

⇒ données de capteurs : l'essor de l'« internet des objets » renvoie au fait que de plus en plus d'objets physiques disposent désormais d'une connexion IP, permettant de générer une grande variété et de gros volumes de données sur leur environnement et leur état. Le potentiel d'objets qui pourraient être connectées d'ici 2020 est estimé entre 30 et 212 milliards selon les sources retenues (Institut Montaigne, 2015).

⇒ données issues des démarches d'open data : on constate que si de plus en plus d'institutions publiques tendent à mettre à la disposition du public les données dont elles disposent, cela est loin d'être le cas du côté des entreprises, et en particulier des firmes issues du numérique puisque la maitrise des données est au cœur de leur modèle économique (Ibekwe-Sanjuan, 2014).

⇒ archives numérisées : l'amélioration constante des procédures de reconnaissance automatique de texte (Optical Character Recognition) fait

qu'il est désormais possible de transformer des sommes d'archives en autant de fichiers lisibles par un logiciel de traitement de texte.

⇒ Donnes issues de questionnaires : la diffusion d'internet a multiplié la possibilité de passation de questionnaires en ligne.

Ensuite, en dépit de leur précision, les données récoltées via les techniques d'enregistrement automatique s'avèrent plus pauvres que prévues (Ollion et Boelaert, 2015). Produites à des fins autres que la recherche, les données de capteurs ne mesurent souvent qu'une partie de ce qui intéresse les chercheurs en général. Par exemple, les informations sur le suivi d'un dossier client (la dernière connexion, la dernière commande, l'accessibilité d'un historique) ne forment pas nécessairement des variables intéressantes pour le scientifique. En d'autres termes, big data ne veut pas nécessairement dire rich data. Ainsi, l'information peut s'avérer plus limitée que si le chercheur avait eu recours à une enquête par questionnaire à la taille limitée. Par ailleurs, la réappropriation de ces données demande souvent, outre leur extraction et leur nettoyage, des procédures de recodage aussi longues qu'appuyées par des connaissances précises du sujet étudié.

Dominique Boullier souligne quant à lui que, en dépit de leur volume massif, les données issues du big data ne peuvent prétendre pour autant englober l'exhaustivité du monde social. De même, malgré leur grande variété, les données numériques ne permettent pas de satisfaire le critère de représentativité qu'exigent les sciences sociales. Autrement dit, les sciences sociales doivent accepter le fait que les traces numériques ne permettent d'éclairer que des questions précises, pour lesquels des traces spécifiques sont mobilisées.

Une autre mise en garde formulée par Ollion et Boelaert concerne le risque de « quantophrénie », à savoir l'appétence immodérée pour les chiffres, indépendamment de ce qu'ils peuvent nous apprendre. L'augmentation massive des volumes de données disponibles pour la recherche pourrait conduire les chercheurs à délaisser les domaines et objets de recherche moins riches en données numériques. Cette tendance pourrait s'avérer préjudiciable dès lors que des pans entiers de la vie sociale ne sont pas traçables via les données numériques.

Au total, ces différentes limites, viennent contredire l'ambition de développer des

outils qui informeraient quasi-immédiatement sur le monde social.

Les algorithmes ne peuvent suffire à donner du sens aux données

Si les données sont toujours construites par une opération, elles doivent aussi toujours être interprétées. Étienne Ollion et Julien Boelaert estiment illusoire de considérer qu'une avalanche de données pourrait en soi améliorer la connaissance. Ainsi, la découverte de structures sociales cachées – « si elles existent » – demande bien plus que de vastes ensembles de données et des outils de recherche de corrélation automatisée. Elle demande aussi une connaissance de première main du sujet et un cadre théorique certes flexible mais déjà établi.

A cet égard, si les méthodes d'apprentissage statistique excellent dans la prédiction et la généralisation et si leur efficacité empirique est largement démontrée, leur grande complexité mathématique fait que cette efficacité échappe encore à toute explication formelle rigoureuse (Ollion et Boelaert, 2015). En bref, les techniques du big data peuvent nous livrer le « quoi » ou le « combien » sans pouvoir nous

fournir les moyens de comprendre le « pourquoi » ni le « comment » des phénomènes observés. Dès lors, selon la chercheuse en science de l'information et de la communication Fidelia Ibekwe-Sanjuan (2014), si le paradigme des big data devait s'imposer nous entrerions dans une science sans but, sans causalité et sans sujets connaissant, dans une rationalité post-moderne fondée sur le calcul de corrélations.

Prolongeant la critique sur le plan politique, l'historien des sciences Evgeny Morozov[3] s'inquiètent que le « techno-solutionisme » qu'incarne le big data puisse conduire à réduire la légitimité de l'action publique à une mesure d'efficacité opérationnelle, évacuant le débat sur les valeurs et les finalités de notre système politique. De même, selon lui, en livrant les décisions qui nous concernent à une optimisation algorithmique, le big data ferait peu de cas du respect des libertés individuelles et de la part de jugement éthique inhérente à tout choix public.

[3] *Cité par Grosdhomme Lulin, 2015*

4. Extension : les « traces » numériques comme nouvelle fenêtre sur le monde social

Selon Dominique Boullier (2015), il revient aux sciences sociales d'apporter une réponse scientifique à la mutation des méthodes de quantification induite par la puissance de calcul et les nouvelles données fournies par le numérique. A ses yeux, cela implique au préalable d'appréhender la révolution numérique en tant que nouvelle réflexivité offerte aux sociétés : de nouvelles sources de données sont désormais disponibles, au-delà des recensements et des registres ou des sondages et des questionnaires. L'enjeu consiste à faire émerger une troisième génération de sciences sociales (voir tableau ci-dessous) qui se caractériserait par l'utilisation de la propagation des traces numériques comme nouveau matériau permettant de suivre une dimension du social jusqu'ici difficile à saisir, « à savoir les vibrations à haute fréquence et non plus les structures sociales de longue durée ni les mouvements d'opinion de moyenne fréquence. » C'est toute l'ambition de la « théorie des vibrations » proposée par Dominique Boullier : bâtir les conventions pour une nouvelle « strate » de sciences sociales portant sur des processus et des entités jusqu'ici incalculables. Comme le résume Dominique Boullier, à chaque longueur d'onde sociale ses méthodes et ses limites de validité : les sciences sociales de troisième génération peuvent à la fois aider à rendre compte de phénomènes inédits et faire préciser le domaine de validité de chaque génération.

Tableau 1. Les trois générations de sciences sociales

	1re génération	2e génération	3e génération
Concept du social	*Société(s)*	*Opinion(s)*	*Vibration(s)*
Dispositifs de collecte	Recensement	Sondage	Traces (*big data*)
Principe de validation	Exhaustivité	Représentativité	Traçabilité
Coconstruction institutions/recherche	Registre/enquête	Audience/sondage	Suivi des traces/analyse des vibrations
Acteurs majeurs de référence (et financeurs)	États	Médias de masse	Marques
Acteurs opérationnels	Instituts nationaux	Instituts de sondage	Plateformes du *web* (GAFAT)
Auteurs fondateurs	Émile Durkheim	George H. Gallup, Paul Lazarsfeld	Michel Callon, Bruno Latour, John Law
Problèmes clés des approches scientifiques initiales	Division du travail et État providence	Propagande et influence des médias (mesures d'audience)	Science et technologie (scientométrie)
Conjoncture technique	Machines de Hollerith (calcul mécanographique)	Radio et téléphone	Internet, *web* et *big data*
Formats sémiotiques	Tableaux croisés et cartes topographiques	Courbes et histogrammes/ diagrammes circulaires	Graphes, *timelines** et *dashboards*
Métriques	Statistique	Échantillonnage	Topologie et *tweet per second* (*TPS*) (Scores)
Critères techniques de qualité des données	Pertinence, précision, actualité, accessibilité, comparabilité, cohérence	Intervalle de confiance, probabilités	Volume, variété et vélocité (*big data*)
Modalités dominantes de la science sociale	Explications	Corrélations descriptives puis prédictives	Corrélations prédictives

Circonscrire la pertinence scientifique des « traces numériques » : on ne peut en tirer des leçons sur la société ou sur l'opinion

Un premier pas essentiel selon Dominique Boullier est de définir le statut scientifique des traces numériques. Selon lui, celles-ci peuvent aller de signaux (« bruts », générés par des objets) à des verbatims (issus de commentaires, posts de réseaux sociaux, avis de consommateurs, etc.), elles peuvent être des métadonnées (plus que les contenus d'un tweet, ces métadonnées sont très riches et aisément calculables), des traces (liens, clics, likes, cookies) exploitées en bases de données par les opérateurs ou les plateformes. Cela les distingue des données que l'on peut récupérer en masse sur des fichiers clients ou encore à partir d'actes administratifs.

Ce faisant, Dominique Boullier conteste l'idée que les traces numériques collectées en ligne permettraient d'accéder au social « véritable » mieux que tous les sondages, toutes les enquêtes et tous les recensements. Selon lui, rien ne permet de garantir quelque lien que ce soit entre les identités de Facebook et des personnes identifiables par l'état-civil et comptées par le recensement. Ce qui est connecté ne sont que des comptes et les données récupérées ne sont que des traces d'activité d'une entité qui peut correspondre éventuellement à des personnes au sens de l'état-civil. Autrement dit, il parait vain à ses yeux de soutenir que « derrière » les sites ou « derrière » les clics, il y a bien des humains équipés de toutes leurs « intentions » (que l'on peut rarement vérifier puisque ce sont des données déclaratives) et de toutes leurs « propriétés sociales » (dont on ne pourra jamais garantir la pertinence pour expliquer un comportement spécifique dans une situation précise).

Selon Dominique Boullier ce constat parait d'autant plus significatif que toutes ces traces, qui pouvaient encore être connectées à des données personnelles, ne seront probablement plus accessibles selon les mêmes conditions dans quelques années. Le succès d'Adblock qui empêche les cookies et autres publicités intrusives, s'amplifie constamment (200 millions de téléchargements, 40 % d'installation sur Firefox en 2014). Le cryptage généralisé deviendra une nécessité face à l'incapacité des plateformes et des services de renseignement à réguler leurs propres activités prédatrices de données personnelles.

Au total, la collecte de ces traces, à la surface des réseaux et sans lien avec les données structurées et socio-démographiquement significatives, ne peut constituer une base solide pour appliquer les modèles de la société et de l'opinion des sciences sociales. Selon Dominique Boullier, les sciences sociales doivent reconnaitre qu'il est impossible d'exploiter les traces numériques pour en tirer de quelconques leçons sur la société ou sur l'opinion, et qu'il est préférable de ne plus parler de personnes mais de comptes, ni de communautés mais de clusters, ni de sociabilité mais de connectivité, ni d'opinions mais de verbatims.

Des traces aux vibrations : suivre à la trace les « vibrations » par lesquelles se produit le changement social

En revanche, Dominique Boullier soutient que les traces numériques permettent de révéler d'autres dimensions du social. D'une part, leur vélocité (collecte en temps réel) modifie le statut des bases de données qui deviennent dynamiques, ce qui apporte une information inédite aux sciences sociales qui, jusqu'ici, étaient d'abord occupées à montrer la force d'imposition de « la société » sur la diversité des comportements individuels à un instant t, ou comment l'opinion publique se structurait au-delà des expressions singulières obtenues dans les enquêtes. D'autre part, les traces portent sur des entités de référence qui ne sont plus des individus, des groupes, la société ou l'opinion.

Selon Dominique Boullier, les traces numériques ouvrent ainsi la voie à une approche centrée sur la traçabilité de masse d'entités détachées de ces catégories mais qui sont constitutives elles-aussi du social : les « vibrations ». Par ce terme, Dominique Boullier évoque les phénomènes de la propagation à haute fréquence d'entités (les traces) dans différents milieux numériques, et dont les propriétés sont susceptibles de créer de petites différences et, par là, d'affecter les individus et les groupes, la société et l'opinion. En d'autres termes, l'analyse des traces numériques rend possible l'analyse de phénomènes parfois anciens (manifestations, modes, olas dans les stades, rumeurs, etc.), mais dont il était impossible de suivre la trace avec les dispositifs habituels des sciences sociales. Cette activité à haute fréquence et à propagation rapide était de fait délaissée ou réduite.

L'exemple de la manifestation du 11 janvier 2015 (je suis Charlie)

Selon Dominique Boullier, la controverse qui a suivi la publication de l'ouvrage[4] d'Emmanuel Todd en mai 2015 – ce livre s'interroge sur le sens à donner à la manifestation du 11 janvier 2015 faisant suite aux attentats de Charlie Hebdo – fut particulièrement significative du malentendu entre les trois générations de sciences sociales et du risque de rapporter toutes les explications à une seule de ces approches. E. Todd a ainsi adopté une posture de « longue durée », renvoyant à l'ancrage religieux des territoires, mobilisant une « mémoire des lieux », qui s'exprime sous forme de « catholicisme zombie » et qui « se perpétuerait alors même qu'elle n'existerait plus qu'à l'état de traces, ou plus du tout en tant que croyance individuelle » (p. 181). Sans discuter ici la thèse elle-même, il apparaît que la manifestation comme phénomène conjoncturel se trouvait ensevelie sous des causes puissantes et intraçables, mais révélées par la vertu totalisante des statistiques historiques. Les spécialistes de l'opinion eurent alors beau jeu de montrer que des sondages avaient été faits après le 11 Janvier et que toutes les données infirmaient les « opinions » attribuées aux manifestants par E. Todd. Les motivations, le sens de la pratique pouvaient être récupérés par la méthode établie des sondages. Ce faisant, la viralité du processus dans la rue comme sur les réseaux numériques n'était pas « expliquée », car ce n'était pas l'objet de ces sondages. C'est ici que des méthodes numériques capables de traiter la vélocité des traces permettraient de rendre compte de la haute fréquence du social sans pour autant invalider les thèses sur la longue durée des appartenances et des croyances ou celles sur la moyenne durée des opinions.

A travers l'analyse de ces vibrations, la troisième génération de sciences sociales vise à mettre au jour non pas la substance de l'opinion de supposés individus mais le pouvoir de circulation d'une vibration se transformant selon les milieux qu'elle affecte. De même, selon Dominique Boullier, l'objectif n'est pas de tendre vers la physique sociale, qui cherche des lois supposées transversales à tous ces flux, car les vibrations sociales gardent leurs particularités et ne s'étudient pas en généralité mais selon les problèmes qui déclenchent leur mise en mouvement (« je suis Charlie » n'est pas de même nature que le « bijoutier de Nice » et aucune « loi » générale ne sera pertinente pour en rendre compte).

Pour Dominique Boullier, l'approche par les vibrations permet de construire une combinatoire infinie, en suivant les extensions, les propagations, les répétitions, à condition de rester centrée sur les problématiques que portent les vibrations. Par exemple, selon lui, il n'est pas suffisant de repérer les clusters que produit la propagation de « je suis Charlie » à des comptes Facebook ou à des sites web pour retrouver d'éventuelles « tendances d'opinion » sous-jacentes (les blogs de gauche, d'extrême gauche, écologistes, etc.). Il faut avant tout restituer la dynamique temporelle et repérer les moments où « je suis Charlie » se transforme visuellement, change de support après Twitter, mute en « je ne suis pas Charlie », ou agrège avec lui un appel à manifester.

[4] *Qui est Charlie ? : Sociologie d'une crise religieuse, éditions du Seuil, 2015*

Un enjeu pour l'avenir : l'accès aux traces numériques

Dominique Boullier alerte cependant sur le fait que la construction de cette nouvelle génération de sciences sociales risque de se heurter aux plateformes du web (GAFAT) qui tendent à occuper tout le terrain de la production et de l'analyse des traces numériques. Ces dernières constituent pour ces firmes une matière première particulièrement profitable, dont l'exploitation s'étend à des domaines d'applications toujours plus larges et vient remettre en question la portée des sciences sociales (end of theory) en matière d'aide à la décision. Dès lors, il parait essentiel selon Dominique Boullier de créer les conditions permettant de fonder les sciences sociales de troisième génération sur une proposition non captive des utilisations qui sont faites de ces traces par les plateformes du numérique, de la même façon que les sondages ne servent pas qu'aux médias ni les recensements aux États.

Ressources documentaires

Chris Anderson, « The end of theory : the data deluge makes the scientific method obsolete », wired.com, 23/06/08

Dominique Boullier, « Les sciences sociales face aux traces du big data. Société, opinion ou vibrations ? », Revue française de science politique 2015/5 (Vol. 65), p. 805-828.

Pierre Delort, « Le big data », Presses Universitaires de France, 2015

Jean-Gabriel Ganascia, « Les big data dans les humanités », Critique 2015/8 (n° 819-820),

Elisabeth Grosdhomme Lulin, « Gouverner à l'ère du Big Data. Promesses et périls de l'action publique algorithmique », Institut de l'entreprise, mai 2015

Hubert Guillaud, « Big Data : vers l'ingénierie sociale ? », InternetActu.net, 20/5/2014

Fidelia Ibekwe-Sanjuan, « Big Data, Big machines, Big Science : vers une société sans sujet et sans causalité ? », XIXème Congré de la Sfsic. Penser les techniques et les technologies : Apports des Sciences de l'Information et de la Communication et perspectives de recherches., Jun 2014

Viktor Mayer-Schönberger, « La révolution Big Data », Politique étrangère 2014/4

Alex Pentland, « Social Physics: How Good Ideas Spread-The Lessons from a New Science », Penguin Press, 2014

Étienne Ollion, Julien Boelaert, « Au delà des big data. Les sciences sociales et la multiplication des données numériques », Sociologie 2015/3 (Vol. 6)

Institut Montaigne, « Big data et objets connectés. Faire de la France un champion de la révolution numérique », avril 2015