15.52 – 3e version
Revue Française de Science Politique, vol. 65, n°5-6, Oct- Déc 2015, pp. 805-828.

Les sciences sociales face aux traces du big data : société, opinion ou vibrations ?

Les sciences sociales dans leur ensemble et la sociologie en particulier ont depuis plusieurs années adapté leurs méthodes et leurs techniques à la puissance de calcul et aux nouvelles données fournies par le numérique. De grands équipements informatiques et statistiques se sont installés dans tous les pays à vocation d'archives (HAL-SHS[1]), de calcul (Progedo[2], Réseau Quételet[3], Cessda[4]), ou de publication (Cleo[5], Hypothèses[6]) pour ne citer que les exemples français. Dans certains cas, les traditions plus anciennes s'équipent de façon plus efficace[7], dans d'autres cas, de nouveaux domaines ou sous-disciplines tentent de faire leur place (enquêtes en ligne[8], bases de données, etc.), comme les *web studies*, au sein d'ensembles plus vastes comme les « humanités numériques »[9]. Dans un cas, le numérique est traité comme une technique de plus, dans l'autre il est traité comme un domaine à part, mais il n'est pas sûr que ces choix suffisent à rendre compte de ce qui est en jeu dans la mutation numérique en cours. En effet, toutes les sciences sociales fournissent une modalité de réflexivité aux sociétés qui reposent sur un ensemble de dispositifs parfois très longs à constituer mais qui, finissant par faire convention[10], apparaissent comme naturels non seulement aux chercheurs mais aux publics, aux décideurs, etc. Nous voudrions proposer ici une grille de lecture de la mutation numérique en tant que nouvelle réflexivité offerte aux sociétés : de nouvelles sources de données sont désormais disponibles, au-delà des recensements et des registres[11] ou des sondages et des questionnaires[12].

[1] Archive ouverte HAL-SHS (Sciences de l'Homme et de la Société).
[2] PROduction et la GEstion des DOnnées en sciences humaines et sociales.
[3] Le Réseau Quetelet coordonne les activités d'archivage, de documentation et de diffusion des données françaises en sciences humaines et sociales.
[4] Conseil européen des archives de données en sciences sociales, Réseau européen des banques de micro-données pour la recherche en SHS (le réseau Quételet est le membre français du CESSDA).
[5] Centre pour l'Edition Electronique Ouverte, qui gère OpenEditions.
[6] Plateforme de carnets ouverte à l'ensemble des disciplines en sciences humaines et sociales.
[7] C'est le cas des historiens lorsqu'ils numérisent leurs sources, lorsqu'ils constituent de nouveaux types d'archives et de nouvelles méthodes d'enquête où les séries de données gagnent une grande visibilité, comme en témoigne le numéro de la Revue d'Histoire Moderne et Contemporaine de 2011, n°58-4bis. Le débat engagé par Armitage et Guldi sur le retour de la longue durée provoqué par l'augmentation sans précédent de la taille des archives désormais numérisées et par l'usage d'outils comme Google N-grams indique bien que les changements d'échelle de la collecte, par exemple, peuvent affecter le cadre de pensée lui-même. Ce que d'autres contestent ou relativisent dans tout le numéro des Annales de 2015 consacré à « la longue durée en débat ». Armitage David, Guldi Jo, « Le retour de la longue durée : une perspective anglo-américaine », *Annales. Histoire, Sciences Sociales* 2/2015 (70e année), p. 289-318.
[8] Par exemple, l'équipement d'excellence DIME-SHS (Données Infrastructures et Méthodes d'Enquête en Sciences Humaines et Sociales).
[9] Pierre Mounier (éd.), *Read/Write Book 2, Une introduction aux humanités numériques*, Open Edition Press, 2013, 230 p.
[10] Eymard-Duvernay F., Favereau O., Orléan A., Salais R. et Thévenot L. ,"L'économie des conventions ou le temps de la réunification dans les sciences sociales », *Problèmes économiques*, n° 2838, Janvier 2004, La Documentation française, Paris.
[11] Alain Desrosières, *La politique des grands nombres. Histoire de la raison statistique*, Paris, La Découverte, 1993. Emmanuel Didier, *En quoi consiste l'Amérique ? Les statistiques, le New Deal et la démocratie*, Paris, La Découverte, 2009.
[12] Loïc Blondiaux, *La fabrique de l'opinion. Une histoire sociale des sondages*, Paris, Seuil, 1998. Plusieurs remarques prennent appui sur ce travail fondamental pour notre réflexion.

Pour mieux nous orienter dans la prolifération des techniques numériques, il parait utile de restituer une perspective historique à ces différentes « époques » de réflexivité, qui, grâce à un long montage technique, institutionnel et économique, ont chacune finit par constituer des évidences non seulement pour les chercheurs mais aussi pour le public. Le tableau synthétique des trois âges des sciences sociales permet de rendre perceptible la cohérence de l'approche comparative mais oblige dans le même temps à schématiser et à éliminer des spécificités propres à chaque âge. Rappelons de plus que nous ne traitons pas ici des versants dits qualitatifs des méthodes des sciences sociales qui pourraient relever d'une même périodisation mais dont l'influence sur la réflexivité des sociétés fut moindre dans la vie quotidienne ou dans le gouvernement des états, des médias ou des marques[13]. Indiquons enfin comme un principe de précaution que si les deux premières générations ont été largement documentées, la troisième constitue en quelque sorte un pari sur la capacité des sciences sociales à organiser une forme de réponse scientifique à une mutation des méthodes de quantification qui pénètre en profondeur tout le social.

Tableau des trois générations de sciences sociales

	1ere génération	**2nde génération**	**3eme génération**
Concept du social	**Société(s)**	**Opinion(s)**	**Vibration(s)**
Dispositifs de collecte	Recensement	Sondage	Traces (Big Data)
Principe de validation	Exhaustivité	Représentativité	Traçabilité
Co-construction institutions/ recherche	Registre/ enquête	Audience/ sondage	Suivi des traces/ analyse des vibrations
Acteurs majeurs de référence (et financeurs)	Etats	Mass media	Marques
Acteurs opérationnels	Instituts nationaux	Instituts de sondage	Plates-formes du web (GAFA)
Auteurs fondateurs	Durkheim	Gallup Lazarsfeld	Callon Latour 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 (camemberts)	Graphes, timelines et dashboards
Métriques	Statistique	Echantillonnage	Topologie et TPS

[13] Ce texte a bénéficié des commentaires critiques de E. Didier, G. Bowker, B. Latour, N. Mayer, D. Cardon, N. Marres, G. Fouetillou, S. Parasie, R. Rogers, F. Cochoy, M. Wievorka, F. Thibault, Y. Citton et M. Legrand, notamment lors d'un séminaire d'un semestre en 2015 sur le thème des sciences sociales de troisième génération à la Fondation Maison des Sciences de l'Homme. Qu'ils en soient vivement remerciés ainsi que Yves Deloye pour la révision finale du manuscrit. Ce texte a fait l'objet d'une première présentation sommaire au colloque Big Data du Collège de France en juin 2014 (chaire de PM Menger).

			(tweet per second) (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

De nouvelles entités, dont le statut reste incertain, sont ainsi rendues accessibles par le numérique, que ni « la société », et ses propriétés socio-démographiques, ni « l'opinion » ne peuvent englober. Nous voulons ici nous inspirer de ces conventions qui ont constitué la réflexivité des sociétés pour examiner précisément les conditions de félicité d'une nouvelle convention qui associerait nécessairement toutes les parties prenantes évoquées dans ce tableau, puisque, de fait, chaque époque des sciences sociales a fait alliance avec des institutions hors de son champ pour produire la quantification de la société ou de l'opinion. Alertons cependant sur le fait que cette nouvelle génération de sciences sociales aura cependant peu de chances d'exister si l'on ne mesure pas que des acteurs (des plates-formes et des producteurs/ capteurs de traces) tendent à occuper tout le terrain[14]. Pour le dire rapidement, le marketing et les « *computer sciences* » s'approprient et génèrent des outils de suivi de la vie sociale, sous forme de suivi de marques, de réputations, de communautés, de réseaux sociaux, d'opinions, etc. qui peuvent se passer des interprétations et des modèles des sciences sociales qu'ils compensent par une puissance de calcul et une traçabilité inédites, celles du **Big Data**[15]. La vélocité des traces[16] ainsi collectées est la caractéristique principale à prendre en compte qui permet de modifier le rythme même de la vie politique lorsqu'elle finit par se centrer sur les tweets, par exemple. Ce que nous appellerons la politique à haute fréquence (ou « high frequency politics »[17]) emprunte ainsi les traits du **High Frequency Trading**[18] de la finance, pour le meilleur et surtout pour le pire, semble-t-il. Le souci

[14] Rappelons que Burrows et Savage avaient alerté bien avant nous toute la sociologie notamment sur l'émergence de « transactional data », numériques et susceptibles de générer une crise à venir de l'empirisme. Savage M. and Burrows R. (2007) The coming crisis of empirical sociology. *Sociology* 41(5): 885–899. Les auteurs ont d'ailleurs récemment mis à jour leurs questionnements à la lumière du Big Data. Roger Burrows and Mike Savage « After the crisis? Big Data and the methodological challenges of empirical sociology », *Big Data & Society*, April–June 2014: 1–6.

[15] Pour la définition précise de certains termes techniques, nous renvoyons le lecteur au glossaire situé en fin d'article.

[16] Les traces numériques n'ont pas le statut des données classiques des sciences sociales ni même celui des traces des historiens comme Charles Seignobos, qui distinguait classiquement les traces directes (vestiges archéologiques, vêtements…) des traces indirectes (écrits imprimés ou archivés…). Charles Seignobos, *La méthode historique appliquée aux sciences sociales*, Félix Alcan, 1901.

[17] Dominique Boullier, « Plates-formes de réseaux sociaux et répertoires d'action collective » in Najar, S. (ed.), *Les réseaux sociaux sur internet à l'heure des transitions démocratiques*, Paris, Editions Karthala, 2013.

[18] Pour avoir une idée de la vitesse de ces transactions, voici ce qu'en dit A. Laumonier : « Le moteur d'appariement du Nasdaq gère par exemple 1 million de messages par seconde (soit 1 par millionième de seconde) – par message on entend un ordre, une cotation, etc. En une seule journée, ce sont donc plus de 25 milliards de messages qui transitent par leur plateforme, ce qui est considérable. (…) En terme de transmission de l'information entre plateformes de négociation (par exemple entre New York, où sont les principaux marchés d'actions, et Chicago, où se trouvent les marchés des dérivés), la limite ultime qui est celle de la vitesse de la lumière dans le vide (soit 300 mètres par milliseconde) est quasiment atteinte. Le réseau de micro-ondes de McKay Brothers, le leader des transmissions entre New York et Chicago, permet aux ordres boursiers de relier les deux villes en 8,12 millisecondes, soit à 95 % de la vitesse de la lumière. Il est peu probable qu'on puisse

principal reste l'action, la réaction, et non l'analyse ou la compréhension telles que les traditions de la sociologie, de la science politique et des autres sciences sociales les avaient définies. Traces et non plus données, réactivité et non plus réflexivité, le monde numérique se trouve façonné par des principes qui laissent de moins en moins place aux formats habituels d'argumentation des sciences sociales. Ainsi la casuistique, c'est-à-dire l'étude détaillée de cas situés dans leurs contextes selon des méthodes dites parfois de « thick description », comportant des dimensions qualitatives et quantitatives, à visée comparative en particulier, n'a plus de statut dans cet environnement qui doit agréger, standardiser et détacher les données de leurs conditions de production pour faciliter une réaction, si possible automatisée. De même, l'approche hypothético-déductive, issue souvent d'une tradition disciplinaire, et exigeant la construction d'un protocole de collecte et de traitement de données orienté par les hypothèses, n'a guère de sens lorsque les machines sont capables de tester des milliers de corrélations et d'en évaluer la robustesse statistique avant même que toute interprétation soit nécessaire. Il n'y a dès lors aucune raison pour que l'autorité des sciences sociales ne soit pas remise en cause comme le furent toutes les autorités depuis l'avènement du numérique en réseaux. Mais, comme le dit Geof Bowker, ce glissement n'est pas sans perturber tous nos repères : «For those of us brought up learning that correlation is not causation, there's a certain reluctance to examine the possibility that correlation is basically good enough. It is surely the case that we are moving from the knowledge/power nexus portrayed by Foucault to a data/action nexus that does not need to move through theory: All it needs is data together with preferred outcomes."[19]. Nous proposons ici de contribuer à bâtir les conventions nécessaires pour faire advenir des sciences sociales et politiques de troisième génération qui utiliseront la propagation des traces numériques comme nouveau matériau en délimitant plus précisément la dimension du social qu'elles permettent de suivre, à savoir les **vibrations** à haute fréquence et non plus les structures sociales de longue durée[20] ni les mouvements d'opinion de moyenne fréquence.

1/ L'âge du numérique

Ni personnes ni identités ni communautés, les traces sont la matière première

Depuis de nombreuses années, mais de façon étendue avec les réseaux sociaux, les « *computer sciences* » calculent et modélisent le social comme si les traces récoltées permettaient d'accéder aux « vrais » individus mieux que tous les sondages, toutes les enquêtes et tous les recensements[21]. Prenons deux exemples, l'un académique et l'autre commercial :

---

physiquement aller beaucoup plus vite. », « L'accélération de la finance. Entretien avec Alexandre Laumonier » Propos recueillis par Auray N., Bourdeau V., Cottin-Marx S., Ouardi S., *Mouvements* 3/2014 (n° 79), p. 92-99.

19 Geof Bowker, « The Theory/Data Thing. Commentary », *International Journal of Communication* 8, 1795-1799, 2014.

[20] La place de la longue durée dans le travail des historiens dans le contexte du numérique est discutée en détail par Lemercier Claire, « Une histoire sans sciences sociales ?. », *Annales. Histoire, Sciences Sociales* 2/2015 (70e année) , p. 345-357.

21 Par exemple, l'une des premières études de masse de Twitter basée sur 41 millions de profils et 106 millions de tweets réalisée par les informaticiens du KAIST en Corée du Sud (What is Twitter, a Social Network or a News Media? Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, Proceedings of the 19th International World Wide Web (WWW) Conference, April 26-30, 2010, Raleigh NC (USA). Ou encore le traitement de la campagne présidentielle américaine de 2008 à partir de la circulation de mèmes sur le réseau (Meme-tracking and the Dynamics of the News Cycle, by J. Leskovec, L. Backstrom, J. Kleinberg. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009).

- « *The Web does not just connect machines, it connects people*." (Knight Foundation, 14 September 2008). Voilà ce que déclarait Sir Tim Berners-Lee, fondateur du web en 1991 avec René Caillau, voulant insister ainsi sur le passage à une dimension du réseau non plus technologique (internet) ni documentaire (le WWW), mais sociale.
- Facebook de son côté a réussi le tour de force de rendre « normal » du point de vue des acteurs eux-mêmes, de déclarer son identité véritable, c'est-à-dire celle fournie par l'état-civil, son nom et son prénom, contre la tradition d'anonymat sur le web. La plate-forme prétend ainsi devenir le monde de référence, voire l'état-civil de substitution, en compétition avec Google sur ce plan.

Or, rien ne permet de garantir quelque lien que ce soit entre les identités de Facebook ou « les gens » de Berners-Lee 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 prendre éventuellement les formats de l'état-civil. Dans le cas des scores qui permettent de classer les sites sur un moteur de recherche, la topologie des sites qui les produit ne traite jamais de leurs contenus en tant que tels, mais des liens entrants et des liens sortants qui produisent un rang d'**autorité** ou de **hub**, au sens de la topologie des réseaux[22] et non d'un statut social. Cette première précision doit nous conduire à assumer la particularité de ce monde des traces sur les plates-formes et à considérer qu'il est impossible de les exploiter pour en tirer de quelconques leçons sur la société ou sur l'opinion. Cela devrait obliger toutes les sciences sociales à 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 **verbatim**s issus de commentaires, etc. dans une approche marquée par la « prudence » de l'empirisme radical[23].

Précisons ici d'emblée ce que nous entendons par traces dans le but de les distinguer des données. Les traces peuvent aller de signaux (« bruts », générés par des objets) à des verbatims non structurés qui se propagent sous forme de **mème**s (ou de citations), elles peuvent être des métadonnées (plus que les contenus d'un tweet, ses métadonnées sont très riches et aisément calculables), des traces (liens, clics, likes, cookies)[24] exploitées en bases de données par les opérateurs ou les plates-formes (GAFAT[25]). Elles peuvent aussi être captées indépendamment de cela à travers les API[26] proposées par ces plates-formes et qui ne relèvent pas alors de bases de données relationnelles[27]. Les traces ne sont pas nécessairement préformatées pour un calcul précis, ni dépendantes de l'agrégation que l'on peut appliquer ensuite. Il est aisé de soutenir que, malgré tout, « derrière » les sites ou « derrière » les clics, il y a bien des humains équipés de toutes leurs « intentions » (qu'on peut rarement vérifier puisque ce sont des données déclaratives) et de toutes leurs « propriétés sociales » (dont on ne pourra jamais garantir la pertinence dans une situation précise pour expliquer un comportement spécifique). Il est plus prudent de suivre les **algorithmes** qui, eux, ne s'intéressent qu'à l'une ou l'autre des propriétés selon leurs visées, propriétés non plus

---

[22] Jon Kleinberg, D. Gibson, P. Raghavan, "Inferring Web Communities From Link Topology", In *Proc. of the 9th ACM Conference on Hypertext and Hypermedia (HYPER-98)*, pages 225-234, New York, June 20-24 1998.

23 William James, *Essays in Radical Empiricism*, 1912. (Traduction française : *Essais d'empirisme radical*, Marseille, Agone, 2005. Réédition poche, Flammarion, Champs, octobre 2007).

24 Dominique Cardon a réalisé un inventaire très précis de quatre types de quantification sur le web selon la position du calculateur : il distingue les clics, les liens, les likes et les traces. Cardon D. (2013), Du Lien Au Like Sur Internet. Deux mesures de la réputation. *Communications*, 93, La réputation, (p. 173-186).

[25] Google, Apple, Facebook, Amazon et Twitter auxquelles il faudrait ajouter Weibo, le twitter chinois.

[26] Application Programmable Interfaces, qui permettent à une application de se connecter aux services et données de ces plates-formes de façon limitée mais suffisante pour produire de l'interopérabilité.

[27] Geof Bowker & Leigh Star S., *Sorting Things Out. Classification and its consequences*. Cambridge, MIT Press, 1999.

structurantes mais parfois apparemment très secondaires (un achat précédent, un like sur un post, etc.) et cela sans aucune garantie mais en acceptant des approximations corrigées par apprentissage sur des masses de données. Les traces, entendues en ce sens restreint, sont produites par les plates-formes et les systèmes techniques numériques, mais ne sont pas les « signes » ou les indices d'autre chose qu'elles-mêmes tant que les relations avec d'autres attributs ne sont pas créées et démontrées.

Cela les distingue radicalement des données que l'on peut récupérer en masse sur des fichiers clients ou encore à partir d'actes administratifs pour lesquels seule la qualité du « Volume » propre au Big Data (en référence aux 3V qu'on lui attribue souvent) est prise en compte. Certes, les méthodes de calcul du Big Data peuvent y être appliquées dans les deux cas, mais les traces sont a priori indépendantes des autres attributs, en particulier socio-démographiques qui sont rarement mobilisés dans les corrélations recherchées entre traces. Les relations avec des paramètres plus classiques dans les sciences des données se limitent au temps (un **timestamp** ou horodatage) et au lieu (les tags de géolocalisation), qui permettent de produire des **timelines** (ou lignes de temps) et des cartes qui deviennent des modes de présentation simplifiés des traces. En s'appuyant sur ces références communes (ce qui souligne l'importance stratégique des Google maps), il est possible de produire des corrélations entre tous types de données, en utilisant alors la qualité « Variété » du Big Data (des actions sur un téléphone portable, des achats dans un magasin, des références dans une base de données disposant des adresses, etc.). Mais grâce à cette simplification radicale, où les entités de référence ne sont plus des individus équipés de toutes leurs propriétés socio-démographiques, les traces circulent vite et modifient le statut même des bases de données qui deviennent dynamiques (on parle même parfois de « temps réel »), ce que les sciences sociales n'avaient jamais eu l'occasion de traiter. La « Vélocité », cette troisième propriété du Big Data, apporte une information inédite pour les sciences sociales et c'est elle qui rend possible l'émergence d'une autre approche centrée sur la traçabilité de masse d'entités détachées des individus. Il faut insister dès maintenant sur la différence entre :

- ces traces faites de clics, de likes ou d'autres marqueurs d'attention éphémère et non verbale (comme les QR Codes[28]),
- les commentaires qui permettent des traitements lexicaux (co-occurrences le plus souvent) qui se rapprochent de ce qui est fait dans les études d'opinion sans avoir aucunement le statut des méthodes établies de mesure de l'opinion,
- les liens hypertexte qui sont des marques d'attachement préférentiel et qui peuvent alors être souvent exploités comme le sont les données des analyses de réseaux sociaux ou comme des marqueurs d'autorité telles que la scientométrie les analysait traditionnellement sous forme de citations. Les analyses des topologies du web[29] que l'on trouve fréquemment dans les « **humanités numériques** »[30] s'appuient en effet sur les propriétés sociales identifiables de tous ces nœuds reliés par les arcs que sont les liens hypertexte.

[28] Franck Cochoy, Smolinski, J. et Vayre, J.-S. (2015), "From marketing to 'market-things' and 'market-ITing': Accounting for technicized and digitalized consumption", in Czarniawska, B. (ed.), *A Research Agenda for Management and Organization Studies*, Cheltenham, UK, Edward Elgar (forthcoming).

[29] Dominique Cardon, Camille Roth et Guilhem Fouetillou, « Topographie de la renommée en ligne : un modèle structurel des communautés thématiques du web français et allemand », *Réseaux*, n°148, 2014.

[30] Dominique Boullier, « La nouvelle fabrique des SHS », préface à Bernard Reber et Claire Brossaud (dir.), *Humanités numériques*, Hermès/Lavoisier, 2007. (traduction : « The new manufacturing of SHS » in Reber and Brossaud, *Digital Cognitive Technologies*, Lavoisier, 2010, pp. xv-xix)

Les traces sont produites par des plates-formes pour les marques

Ces traces numériques constituent une « matière première » particulièrement profitable pour les plates-formes. Les méthodes de marketing digital reposent ainsi largement sur de l'adressage de masse de publicités ou de mails à des adresses IP ou mails qui ont cliqué sur un article (*retargeting*) mais beaucoup plus rarement sur des mises en relation sophistiquées avec les autres attributs des supposées personnes attachées à ces adresses ou à ces clics (*profiling*). Les traces sont une des ressources clés pour les marques pour suivre les effets de leurs propres actions sur leur public. La réputation ou la notoriété ne se traduisent plus seulement dans des mesures d'audience[31]. Sur les réseaux, il leur faut mesurer à la fois une forme d'audience (le reach), des activités élémentaires de ranking réalisées par ces publics incertains (likes, étoiles), mais aussi des activités plus élaborées, comme leurs commentaires, qui constituent ce qu'on appelle leur « taux d'engagement ». Les marques sont friandes de ces traces et ce sont elles qui alimentent le chiffre d'affaires de toutes ces plates-formes et par là de tout le web, rappelons-le. Les outils d'opinion mining et de sentiment analysis[32] constituent ainsi la réponse à cette angoisse du marketeur après le lancement de produit. Cependant, l'extension du domaine de la marque atteint toutes les activités sociales, qu'elles soient commerciales, culturelles, politiques, institutionnelles voire interindividuelles lorsque chacun doit mesurer son excellence à l'aide de rankings, comme le font les chercheurs eux-mêmes[33] et cela malgré les critiques vigoureuses et largement partagées de ces indicateurs[34]. Dès lors, ce sont les méthodes des marques qui prennent partout le dessus et imposent leur loi et leur rythme, jusque dans les services publics qui doivent pratiquer le benchmarking. 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 *(features*) de la marque qui sont affectées.

Le monde politique lui-même est désormais pris dans cette spirale de la réactivité et son addiction aux tweets nous a conduit à considérer que nous étions entrés dans l'ère du *High Frequency Politics* à l'image du *High Frequency Trading* de la finance spéculative. Dès lors les puissants phénomènes de viralité propres à la plate-forme Facebook et à son mécanisme de likes, sont spectaculaires et interrogent les chercheurs en science politique[35]. Dans l'affaire du bijoutier de Nice (autodéfense d'un bijoutier qui abat son agresseur), la page a été créée le 11 septembre 2013, elle comptait 1 000 000 de likes le 14 septembre et est allée jusqu'à 1 635 000 likes le 7 Novembre. La propagation des « #JesuisCharlie » s'est faite dans le

[31] Bien que la tradition des mesures d'audience soit directement liée à la construction de l'opinion par les sondages et à l'échantillonnage, on continue à parler de façon abusive de mesures d'audience sur le web sans avoir aucunement les cadres d'analyse de cette tradition, bien analysée par Cécile Méadel, *Quantifier le public. Histoire des mesures d'audience de la radio et de la télévision*, Paris, Economica, 2010.

[32] Dominique Boullier et Audrey Lohard, *Opinion mining et sentiment analysis. Méthodes et outils*, Paris, Open Editions Press, 2012.

[33] Isabelle Bruno, Emmanuel Didier, *Benchmarking. L'Etat sous pression statistique*, Paris, La Découverte, coll. Zones, 2013.

[34] Yves Gingras, *Les dérives de l'évaluation de la recherche. Du bon usage de la bibliométrie*, Paris, Éditions Raisons d'agir, 2014.

35 Pour une revue de ces approches plutôt centrées sur les formes d'expression et de communication, Fabienne Greffet, Le web dans la recherche en science politique. Nouveaux terrains, nouveaux enjeux, Revue de la BNF, 2012/1 (n° 40), pp.78-83, et sur les propriétés socio-démographiques des internautes en politique, Anaïs Theviot, *Qui milite sur Internet ?Esquisse du profil sociologique du « cyber-militant » au PS et à l'UMP*, Revue Française de Science Politique, 2013/3 (vol.63), pp.663-678.

monde entier[36] : les tweets utilisant le hashtag, l'image et le slogan créés par @jachimroncin démarrent dès le 7 Janvier à 12h52 pour atteindre le soir même plus de 6500 Tweets par minute soit 3,4 millions de tweets en 24 heures, 3 pages et groupes Facebook atteignant plus de 400 000 membres. Dans tous ces cas, limités ici à des thèmes propres à intéresser la science politique (car le buzz peut porter sur tout), les effets de contagion sont impressionnants et méritent une approche spécifique. Dans le cas du bijoutier de Nice, cela a suscité des analyses de tous types, des plus constructivistes critiques (« tous les likes sont achetés ») jusqu'aux plus positivistes (« c'est bien la preuve que l'opinion, - ou « les gens » - ont basculé dans le réflexe sécuritaire »). Pour Charlie, il fut souvent rappelé que ce public utilisateur de la plate-forme Twitter n'était pas représentatif de toute la population, ce qui est incontestable sans pour autant que cela enlève toutes ses qualités à ce matériau nouveau. Pire, on peut se demander légitimement si le monde Twitter, par exemple, ne tourne pas autour de lui-même mesurant son rôle au nombre de Retweets à l'exclusion de toute autre influence. Cette tendance autoréférentielle est déjà pointée par des spécialistes du marketing pour calmer l'excitation générale des communicants en faveur de ces réseaux alors que les « taux de conversion » (ventes effectives) sont parfois non mesurables ou décorrélés. Pour ces raisons au moins, il est légitime que la science politique et les sciences sociales en général s'interrogent sur leur usage de ces sources. Plusieurs postures sont possibles et les conventions de recherche à construire ne sont pas du tout identiques.

Faire les mêmes sciences sociales à partir de ces nouvelles sources de données

Le montage des conventions associées aux traces est encore en cours et la prolifération des services d'opinion mining que nous avons étudiée[37] de même que leur extrême diversité en qualité en est la preuve. De nombreuses recherches en sciences sociales continuent à travailler avec les paradigmes existants et exploitent les données collectées sur les réseaux numériques pour les traiter selon leurs méthodes habituelles, tout en apportant des corrections et des contrôles nécessaires. Nous ne parlons pas ici des enquêtes en ligne ou des traitements de masse de corpus textuels bien balisés tels que des corpus d'articles scientifiques que traite la scientométrie, car ces approches n'exploitent pas de traces nativement en ligne ni produites par les plates-formes. Les *web studies* issues des sciences sociales mobilisent, elles, les mêmes cadres classiques de la sociologie appliquées à ces nouvelles sources de données: études économiques de préférence à partir des requêtes Google, études de sociabilité en réseaux, suivi longitudinaux de « communautés », autour de thèmes, ou de sites particuliers, approches d' « *opinion mining* » et de « *sentiment analysis* » mobilisées pour augmenter le suivi de l'opinion publique ou le repérage de tendances. Les graphes réalisés à partir des réseaux sociaux ou des liens entre sites pour produire des communautés d'intérêt ou des cartes sémantiques d'un domaine controversé ne changent pas de paradigme mais l'équipent avec de nouveaux outils. Le numérique **amplifie**[38] ainsi « la réalité de l'opinion » ou confirme les « structures sociales », objets constitués dans les traditions académiques bien avant le numérique.

[36] Louise Merzeau, « #jesuischarlie ou le médium identité », *Médium*, 43, avril 2015, pp. 36-46.
[37] Boullier et Lohard, 2012, op. cit.
[38] Elisabeth L. Eisenstein, *La révolution de l'imprimé dans l'Europe des premiers temps modernes*, Paris, La Découverte, 1991.

Des auteurs comme Monroe et al.[39] ont donné plusieurs exemples de ce qui pourrait être traité du point de vue de la science politique par des méthodes de Big Data en s'intéressant aux phénomènes observables en ligne. Ainsi ces auteurs rendent compte d'une étude[40] conduite sur les réseaux sociaux chinois qui a permis d'observer que les publications sanctionnées par la censure n'étaient pas celles qui critiquaient le régime mais celles qui tentaient de lancer des actions collectives[41]. Pour vérifier cette hypothèse, les chercheurs ont produit eux-mêmes des publications selon ces deux critères et ont pu vérifier la validité de leurs hypothèses, après avoir réalisé une forme de sociologie expérimentale. D'autres[42] ont exploité les traces laissées par les requêtes (Google Search) lors de la campagne Obama pour comparer les proportions de termes racistes dans les requêtes selon les états américains. Ils ont considéré que les requêtes constituaient des expressions spontanées qu'ils n'auraient jamais obtenues par des méthodes conventionnelles d'entretiens par exemple.

Pratiquer les méthodes numériques en détournant les traces

Une autre approche est possible qui permet de récupérer les traces nativement numériques produites par les plates-formes en les détournant pour des usages scientifiques. R. Rogers propose ainsi de faire du « repurposing »[43] de ces traces en travaillant par exemple le « query design », la formulation des requêtes sur Google pour obtenir des réponses à des hypothèses bien construites. Wikipédia constitue un cas à part fort précieux pour les chercheurs car, ainsi que le montre le projet Contropedia, la plate-forme garde l'historique de toutes les interventions, de tous les débats et controverses et produit des arbitrages dont les traces sont aussi gardées. R. Rogers a ainsi étudié les polarisations politiques et les termes des controverses sur l'avortement ou sur le massacre de Srebenica selon les versions croate, bosniaque ou serbe. Mais Wikipédia est unique sur ce plan. Rogers a aussi exploité Twitter comme une « narrative machine » pour reconstituer la révolte iranienne de 2009 à l'aide des 600000 tweets qu'il a pu constituer en corpus. Dans ce cas, les effets de propagation sont bien présents, comme dans le cas du slogan « Dégage ! » des printemps arabes. N. Marres[44] est plus soucieuse de répondre aux critiques sur la dépendance des chercheurs aux données fournies par les plates-formes. Son approche des controverses socio-politiques par les « *issues* » constitue un parti-pris très fort qui permet de fournir des limites de validité aux recherches empiriques (pas de suivi général de tweets- ou de weibos- ou d'autres traces sans arène constituée par les « issues »). Elle estime qu'il ne faut pas chercher à purifier les paramètres des médias car ils font partie du process. Mais il est possible de les corriger comme elle le fait par exemple pour suivre les controverses sur le DPI (Deep Packet Inspection) dans un congrès WCIT (World Conference on International Telecommunications) en 2012 à Dubaï. N. Marres a croisé les hashtags les plus fréquents sur Twitter avec des termes utilisés par les experts dans leurs documents pour refaire les requêtes sur un corpus

[39] Burt L. Monroe, J. Pan, M. Roberts, M. Sen, B. Sinclair, No ! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science, *American Political Science Association*, January 2015, 71-74. Merci à N. Mayer pour le signalement de cet article.

[40] Gary King, J. Pan and M.Roberts, « How Censorship in China Allows Government Criticism but Silences Collective Expression », *American Political Review*, 107 (2), 2013, pp. 326-343.

[41] Cette approche fait écho aux travaux plus qualitatifs de Séverine Arsène sur les cadrages de leurs propres prises de parole réalisés par les internautes chinois dans un contexte de censure : Séverine Arsène, « De l'autocensure aux mobilisations. Prendre la parole en ligne en contexte autoritaire», *Revue française de science politique* 5/2011 (Vol. 61), p. 893-915.

[42] Seth Stephens-Davidowitz, « The Cost of Racial Animus on a Black Presidential Candidate : Evidence Using Google Search Data », *Journal of Public Economics*, 118, 2014, pp.26-40.

[43] Richard Rogers, *Digital Methods*, Cambridge, MA, MIT Press, 2013.

[44] Marres, Noortje, « Why Map Issues ? On Controversy Analysis as a Digital Method », *Science, Technology and Human Values*, March 2015, pp.1-32.

Twitter de façon plus pertinente et indépendante des calculs de la plate-forme. Cela donne une idée de la façon dont il convient de prendre en compte de façon très détaillée et méthodique les effets de cadrage produits par les médias, comme le font d'ailleurs les chercheurs avec les conditions de réalisation des sondages ou des enquêtes non numériques[45].

Construire les conventions des sciences sociales de troisième génération

Nous proposons ici d'adopter un point de vue d'empirisme plus radical[46] que les approches précédentes en considérant que les traces numériques produites par les plates-formes (actuelles et à venir) doivent être traitées dans leur milieu écologique et non rapportées à des mouvements de société ou d'opinion. Cette auto-restriction prend au sérieux les critiques émises sur les « biais » produits par ces plates-formes mais les assument, en considérant que ces traces numériques rendent compte d'autres phénomènes, mobilisent d'autres entités que celles des sciences sociales de la société ou de l'opinion. Les phénomènes de propagation à haute fréquence d'entités (les traces) dans différents milieux numériques permettent de suivre et de calculer enfin les processus d'imitation (invention et opposition en même temps) que Gabriel Tarde avaient identifiés jadis[47]. Cela ne remet pas en cause les autres analyses sur les structures sociales de longue durée et les mouvements d'opinion de moyenne durée mais aucune de ces deux approches ne doit réduire à ses propres principes ce que nous appelons des « vibrations » qui sont aussi constitutives du social. Le pari n'est donc pas de faire une extension des terrains des sciences sociales classiques (correspondant aux deux premiers âges des sciences sociales évoqués au début de cet article) ni une spécialité de méthodes traitant de phénomènes identiques mais bien de produire les conventions pour une nouvelle « strate » de sciences sociales portant sur des processus et des entités jusqu'ici incalculables. Mais il convient alors de veiller à maintenir les ambitions de réflexivité et de critique des sciences sociales face aux tendances à la réactivité pure des agences et des plates-formes qui utilisent les mêmes données. Le calcul est rendu d'autant plus possible qu'il porte sur des entités simples, élémentaires et quasi similaires. « La statistique est un dénombrement d'actions similaires, le plus similaires qu'il se peut », disait G. Tarde.
Pour construire ces conventions, il nous faut entrer un peu plus en détail sur certaines conditions de félicité en partant de ce que fait le Big Data avec toutes ces traces. Les tendances du Big Data peuvent en effet fournir des premières pistes de ces conventions qui méritent d'être confrontées à celles des sciences sociales précédentes. Ainsi les critères de qualité du Big Data sont souvent résumés aux 3V déjà mentionnés: Volume, Variété, Vélocité. La parenté avec les exigences des sciences sociales est ici assez frappante, ce qui justifie notre démarche.

Volume et exhaustivité
Le volume correspond à l'exigence d'exhaustivité traduite sous un mode quelque peu limité, puisque personne ni rien ne permettent de définir les frontières des univers de données

[45] Nonna Mayer, *Sociologie des comportements politiques*, Paris, Armand Colin, 2010. Les questions du rôle des plates-formes dans le cadrage des expressions peuvent très bien se lire dans la tradition des travaux sur *l'agenda-setting* (Maxwell McCombs et Donald Shaw, « The Agenda -Setting Function of Mass Media », *Public Opinion Quaterly*, 36 (2), pp176-187) et sur le *framing* (Iyengar, S., *Is Anyone Responsible ? How Television Frames Political Issues*, Chicago, University of Chicago, 1991).
[46] William James, *Essays in Radical Empiricism*, 1912. Traduction française : *Essais d'empirisme radical*, Agone, Marseille, 2005. Réédition poche, Flammarion, Champs, octobre 2007
[47] Gabriel Tarde, *Les lois de l'imitation*, Paris, Félix Alcan, 1890.

rassemblées. L'absence de « tout » de référence sur le web[48], dans un système dynamique à haute fréquence, ne permet pas de construire clairement un « univers » pour produire des statistiques classiques. Nous devons donc faire notre deuil de l'exhaustivité sans pour autant abandonner l'impératif de préciser les conditions minimales d'acceptabilité d'un corpus donné de traces du point de vue du volume.

Variété et représentativité

Le second critère, la variété, constitue une forme de transcription de cette exigence de représentativité qui a permis à toutes les sciences sociales de procéder par enquêtes, par sondages, à base d'échantillonnage. Cependant, le critère « variété » est une version lâche de la représentativité, puisqu'il se contente d'accepter un niveau *suffisant* de variété. Pour les sciences sociales de troisième génération qui acceptent de perdre la contrainte de représentativité telle qu'elle a été construite dans le cas des sondages, il reste à définir ce que serait cette variété. La constitution d'un ensemble de sources (*sourcing*) lors d'études du web par exemple devrait alors répondre à quelques critères propres aux méthodes numériques et au domaine étudié. Nos travaux sur l'opinion mining nous ont conduit à considérer qu'aucune description du social-société, du social-opinion ou du social-traces ne peut être produite en généralité sur les réseaux numériques. La prolifération des traces rend impossible toute prétention à une référence à un « tout » posé a priori ou constitué a posteriori. Les sciences sociales doivent accepter de ne traiter que des « *issues* »[49], ou des points de focalisation d'attention, dont le numérique peut garder les traces, des traces qui seront spécifiques à chaque *issue*. Cela réduit considérablement la portée totalisante des prétentions du Big Data mais cela rend possible une certaine traduction des impératifs de représentativité et d'exhaustivité.

Vélocité et traçabilité

Enfin, le dernier critère, la vélocité ne trouve guère d'équivalent dans les sciences sociales jusqu'à présent. A vrai dire, ces processus dynamiques ne constituaient pas leur point fort ni leur préoccupation. Il était en effet essentiel de trouver avant tout à représenter les positions à un instant *t*, pour montrer la force d'imposition de « la société » sur la diversité des comportements individuels ou pour montrer comment l'opinion publique se structurait au-delà des expressions singulières obtenues dans les enquêtes. Certes, à travers un suivi longitudinal très coûteux des mêmes populations ou la reprise des mêmes questionnaires, il était possible de restituer un équivalent d'une dynamique, sans jamais cependant pouvoir suivre à la trace les médiations qui auraient permis de produire ces évolutions. La vélocité semble donc hors du champ des approches plus classiques.[50] Or, ce critère nous parait le socle sur lequel faire émerger l'analyse de phénomènes parfois anciens (pensons aux manifestations, aux modes, aux olas dans les stades, aux rumeurs, etc.) mais dont il était impossible de suivre la trace avec les dispositifs classiques des sciences sociales. Cette activité à haute fréquence et à propagation rapide était de fait délaissée ou réduite, comme l'a fait E. Todd pour la manifestation du 11 Janvier 2015[51] (#jesuisCharlie), à une « hystérie » entièrement déterminée par des causes d'autant plus puissantes qu'elles sont invisibles et

[48] Bruno Latour, Jensen P., Venturini T., Grauwin S., Boullier D., « The Whole is Always Smaller Than Its Parts'. A Digital Test of Gabriel Tarde's monads », *British Journal of Sociology*, 2012, Volume 63, Issue 4, pages 590–615.

[49] Noortje Marres, "The Issues Deserve More Credit: Pragmatist Contributions to the Study of Public Involvement in Controversy", *Social Studies of Science*, 37, 2007, p. 759-78. Noortje Marres and Weltevrede, E., "Scraping the Social? Issues in live social research", *Journal of Cultural Economy*, 6 (3), 2013, p. 313-335

[50] Nous ne mentionnons pas ici les approches de la physique sociale, de l'éthologie ou de l'épidémiologie sociale qui ont produit des modèles sociaux sans référence aux traditions des sciences sociales.

[51] Emmanuel Todd, *Qui est Charlie ? Sociologie d'une crise religieuse*, Paris, Seuil, 2015.

lointaines (et donc non documentables), selon le procédé d'Emile Durkheim dans son étude des origines des religions avec son « Dieu-société »[52]. La controverse qui a suivi la publication de l'ouvrage de Todd fut particulièrement significative du malentendu entre les trois générations de sciences sociales et du risque de rapporter tous les comptes rendus voire même les explications à une seule de ces approches. Todd a ainsi adopté une posture « 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 apparait que la manifestation comme phénomène conjoncturel se trouvait ainsi ensevelie sous des causes puissantes et intraçables, mais révélées par la vertu totalisante des statistiques historiques. Les spécialistes de l'opinion[53] 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.
Cependant, une branche des sciences du web s'est, elle aussi, emparée de cette question de la vélocité à sa façon en exploitant les traces des mèmes qui se propagent sur le web. Il est très significatif que Kleinberg, celui-là même qui avait exporté les méthodes de la scientométrie vers l'étude de la topologie du web, méthodes qui furent reprises par Google, se soit intéressé depuis plusieurs années[54] (2002) à la mise au point d'un « meme tracker » avec Leskovec[55]. Leur étude la plus fameuse a porté sur la propagation des citations durant la campagne Obama, ce qui leur permit de réaliser une visualisation spectaculaire de la focalisation de l'attention en courbes à montées et descentes très rapides (*streams and cascades*) autour de certains incidents de la campagne[56]. Leur méthode agrège tous les types de traces que peuvent laisser ces citations, traitées comme des chaines de caractères dont on peut trouver la trace dans tout le web. Elle en produit une métrique ancrée dans le temps, au jour le jour, voire minute par minute désormais avec Twitter (l'unité de mesure étant même devenue le Tweet par seconde). La prise en compte des **mèmes** nous parait prometteuse sous réserve que l'on suive aussi les transformations-traductions de ces mèmes dans des milieux différents. Toute onde en effet se propage dans un milieu qui possède des propriétés de réfraction et de diffraction différentes. Dès lors, il est légitime de se demander ce que ces travaux ont apporté par exemple à la science politique puisqu'il s'agissait d'une campagne électorale. Les auteurs insistent avant tout sur leur apport pour la compréhension du « news cycle » entre blogs et médias *(« the dynamics of information propagation between mainstream and social media »*). Formulés ainsi, les résultats semblent agnostiques sur le contenu des mèmes tracés dans ce

[52] Bruno Latour, « Formes élémentaires de la sociologie, formes avancées de la théologie », *Archives de sciences sociales des religions*, 167, juillet-septembre 2014, p. 255-277.
[53] Nonna Mayer et Vincent Tiberj (Le Monde du 19 Mai 2015) et Luc Rouban (notes du Cevipof du 13 Mai 2015)
[54] Jon Kleinberg, "Bursty and Hierarchical Structure in Streams", *Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2002. J. Kleinberg part ici d'une détection de « topics » qui provoquent des « bursts » dans des flux d'emails, son objectif étant d'aider à les structurer et non de les suivre en tant que tels. Mais il fait le lien avec les problèmes du temps dans la narration tels que traités par Genette.
[55] Jan Leskovec J., L. Backstrom, J. Kleinberg, 2009, op. cit.
[56] Notons la parenté des streams avec les « streams of thought » et « streams of consciousness » de W. James dans ses Principes de Psychologie, pour qui « ça pense ». William James, *Précis de psychologie*, Les empêcheurs de tourner en rond, Paris (1e édition : 1892).

cycle et relèvent plus des « media studies ». Pour autant, les auteurs annoncent un programme de travail plus pertinent pour la science politique : « *one could combine the approaches here with information about the political orientations of the different news media and blog sources to see how particular threads move within and between opposed groups* ». Cette proposition indique clairement qu'une exploitation en termes de science politique nécessiterait de mobiliser des catégories classiques de l'opposition politique pour comprendre la relation entre entités circulantes et le milieu qu'elles pénètrent. Mais, même si les recherches annoncées n'ont pas été conduites par les auteurs, on voit d'emblée le risque de tautologie car ces classifications seront difficiles à questionner, alors qu'une approche d'empirisme radical aurait attendu de la courbe de propagation qu'elle donne accès précisément à ces parcours et à des voisinages non classiques. Sortir de l'autoréférence sur les médias pour dire quelque chose des processus politiques fondés dès lors sur des « opinions » (et des camps identifiés) présente dès lors un risque certain si la puissance de connectivité des entités n'est pas préservée face à une supposée « structure » du milieu de propagation.
Une fois ces précautions établies, il devient cependant possible de trouver un équivalent scientifique de la vélocité du Big Data : la ***traçabilité***. Elle devient le critère essentiel de qualité des entités que l'on peut étudier. Quelques conditions de félicité doivent être réunies pour y parvenir :

a/ Les traces en question doivent avoir une *continuité* suffisante pour qu'il soit encore possible de dire qu'il s'agit d'un même processus, sans avoir la contrainte de la première mémétique[57] (identique) ni le laxisme d'une intertextualité généralisée, où tout pourrait être signe et reprise de tout. Les traces les plus simples sont de ce fait les plus aisées à suivre : des likes sur une page Facebook ou des hashtags repris sur Twitter par exemple.

b/ Les traces en question doivent permettre des suivis d'associations hétérogènes, c'est-à-dire une *puissance de connectivité* suffisante. Pour cette raison, des traces dont le format est trop spécifique à une plate-forme peu connue ne peuvent donner lieu à extension ni à suivi.

c/ Le suivi des traces en question doit permettre de *dater* avec précision tous les événements, toutes les transformations et toutes les associations. Les **timelines** sont ici l'équivalent d'autres conventions, comme les points cardinaux pour la topographie ou les niveaux de revenus pour les sciences sociales de première génération.

Ces conditions élémentaires rendent possible le basculement des sciences sociales vers le suivi d'éléments qui ne sont plus ni des individus ni des groupes, ni la société, ni l'opinion. Le numérique en réseaux a amplifié l'incertitude sur le statut de ces catégories, qui avaient été déjà remises en cause dans des approches comme l'ethnométhodologie. La propagation de ces éléments, qu'il nous faudra qualifier au-delà des traces techniquement collectées, devient l'objet d'étude de la troisième génération de sciences sociales car ce sont les propriétés de ces entités qui leur permettent de créer de petites différences et par là de circuler et d'affecter les individus et les groupes, la société et l'opinion.

Cependant, les conditions de faisabilité de cette exploitation des traces doivent tenir compte, comme l'ont montré les approches en méthodes numériques évoquées précédemment, de la dépendance aux plates-formes qui s'installe ainsi. On ne peut guère espérer modifier ces traces à la source, tant le rapport de forces est en faveur des plates-formes. Il est en revanche possible d'exploiter les traces produites par les plates-formes en les *détournant* de l'usage pour lequel elles avaient été conçues (*repurposing*) comme le propose R. Rogers mais avec

57 Richard Dawkins, The Selfish Gene. Oxford, Oxford University Press, 1976. Susan Blackmore, The meme machine, Oxford, Oxford University Press, 1999.

l'objectif d'un suivi d'entités purement numériques. La règle veut ici qu'on ne prenne en compte aucune explication d'un autre niveau ou d'un autre monde non numérique, pour pouvoir comparer des vitesses de propagation, des rythmes, des transformations éventuelles (par exemple par contamination sur d'autres domaines, etc.). Cette nouvelle génération de sciences sociales devra faire la preuve qu'elle fait émerger des processus qui n'avaient jamais encore été mis en évidence, en raison soit des limites techniques pré-numériques, soit des cibles adoptées par les générations précédentes de sciences sociales.
Le champ est largement ouvert pour cette approche supplémentaire du social car l'ère des traces ne fait que commencer et les plates-formes ne sont pas et ne seront pas les seuls fournisseurs de traces en masse. L'internet des objets n'est plus un fantasme d'ingénieur, et la vie ordinaire commence à se peupler d'échanges sans contact, de puces RFID et d'autres géolocalisations automatiques qui dépendent non plus des personnes mais des objets eux-mêmes, l'interobjectivité de B. Latour[58] devenant désormais traçable. Leurs traces durant leur parcours, leur état (activé ou non) permettent de piloter des processus logistiques, transactionnels, qui sont souvent confinés aux mondes professionnels concernés. Cependant leur extension et leur accès ouvert seront quasiment inévitables dès lors qu'on s'engagera dans une prolifération déjà annoncée. Il ne sera plus possible de renvoyer à des « personnes », à des entités « sociales » au sens des sciences sociales classiques. Il n'y a pas de raison pour que les sciences sociales ne s'emparent pas de ces nouvelles sources.

Big Data sans théorie ?

Cependant, cette ambition théorique et politique d'établissement d'une nouvelle génération de sciences sociales pourrait se heurter à la force de frappe des méthodes du Big Data qui ne relèvent pas seulement des trois V évoqués mais aussi d'une approche corrélationniste radicale et non théorique. Chris Anderson du magazine Wired avait lancé un pavé dans la mare avec son court papier intitulé « The End of Theory » en 2008[59]. Sa remarque se fondait sur une observation des pratiques dans plusieurs domaines scientifiques (génomique ou physique) où l'usage massif des données à des fins de découverte de corrélations avait remplacé selon lui la nécessité des modèles, des hypothèses et des épreuves construites pour les tester. Pour lui, de la même façon, le web constituait un autre monde qui ne décalquait rien mais mettait à disposition les outils pour capter les traces, les agréger, les calculer, les évaluer et les utiliser pour les modifier en retour, dans le même mouvement. Les likes n'ont pas besoin de théorie. La plate-forme capte les traces des actions des internautes (ou des machines) qui cliquent, sous un format standardisé, elle les agrège et produit un score, qui est affiché et peut être utilisé par la plate-forme elle-même pour afficher des tendances. Celles-ci permettent d'orienter les placements publicitaires des annonceurs qui, eux aussi, cherchent à mesurer les effets de leurs choix de placements ou de communication. Dans son format réduit, voilà la chaine qui a été produite. La théorie sociale n'a quasiment aucune utilité dans un tel dispositif opérationnel où le mécanisme performatif fonctionne quasiment à l'identique de celui des mesures d'audience. S'il se limitait à ce secteur du marketing, ce phénomène des traces ne serait finalement qu'une extension de la bulle auto-référentielle si typique de la finance devenue spéculation, jeu de miroir permanent ou « économie d'opinion »[60]. La remise en cause de la nécessité de la théorie touche plus profond et provoque le malaise évoqué par G. Bowker dans notre introduction. La quête des causes, souvent finales, n'est pas une qualité

[58] Bruno Latour, « Une sociologie sans objet ? Remarques sur l'interobjectivité », *Sociologie du travail*, 4, 1994, p. 587-607.
59 Chris Anderson, « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », *Wired Magazine*, 23/06/2008.
60 André Orléan, *Le pouvoir de la finance*, Paris, Odile Jacob, 1999.

scientifique en tant que telle, car bien souvent, elle peut s'accommoder d'argumentation empirique très lacunaire, ce que l'on trouve souvent dans les approches critiques, ou de mises à l'épreuve très réduite des catégories utilisées, une caractéristique du positivisme.

Ce qui est nouveau avec le mode de raisonnement associé au Big Data n'est pas tant l'absence de détours par les causes pour rendre compte des données recueillies, que l'absence de questionnement des propriétés et de la validité de ces données formatées par les systèmes informatiques pourtant très puissants. Cela permet de lancer des recherches de corrélation entre données de tous types du seul fait qu'elles sont disponibles. La puissance de calcul est désormais abondante, la collecte et le stockage de données semblent sans limite, et il devient possible de tester toutes les corrélations sans restriction. Chacun sait pourtant que « corrélation n'est pas raison » mais pour autant tous les travaux à base statistique les mobilisent à condition de respecter certaines procédures et de poser certaines hypothèses qui sont alors mises à l'épreuve par le calcul. Dans le cas du Big Data, la procédure est inversée : les hypothèses sont générées par les corrélations dès lors qu'elles sont suffisamment robustes ce qui revient à une forme d' « automatisation de l'induction », qui dépend entièrement des catégories implicitement présentes dans les données. Et les machines finissent en effet par trouver des corrélations parce qu'elles peuvent choisir dans ce qu'on appelle des portfolios d'algorithmes ceux qui sont les plus performants. Non seulement les données peuvent être combinées à volonté mais les algorithmes aussi. Les erreurs ainsi engendrées ne sont pas un problème puisque des méthodes d'apprentissage permettent d'en tirer des leçons au fur et à mesure au-delà des exemples de départ validés par des experts dans une démarche d'apprentissage supervisé.

C'est pourquoi le Big Data n'est pas seulement une affaire de volume mais dépend en grande partie de cette nouvelle version de l'intelligence artificielle qu'est le **Machine Learning**. Des modèles d'apprentissage ont été élaborés[61] et sont tout aussi importants pour comprendre les enjeux du Big Data. Ce principe d'essai/erreur à haute fréquence finit par 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 sans doute pas le travail d'interprétation mais lorsque l'enjeu consiste à agir (à trouver des protocoles thérapeutiques, à générer une activité en ligne sur le nom de marque, par exemple), l'important est le résultat et non son explication. Ces approches exercent alors une véritable attraction pour des décideurs, notamment en situation de responsabilité politique. Les sciences sociales pourraient dès lors faire face à un relatif désinvestissement si le souci de l'action à court terme l'emportait totalement sur celui de l'explication. Ainsi l'évaluation des politiques publiques (et cela dans tous les domaines qui peuvent relever d'une démarche de benchmarking à partir de séries de données suffisamment importantes) peut tout à fait basculer vers ce type d'approche qu'on pourrait dire agnostique sur le plan théorique. A la condition de ne jamais questionner les entités tracées par les systèmes de recueil de données ni leurs conditions de production. C'est en cela qu'on peut parler d'un « positivisme algorithmique » qui ne peut proliférer qu'en raison d'un certain épuisement des modèles explicatifs de grande portée dans les sciences sociales. La théorie des vibrations prétend contribuer à renouveler cette ambition théorique tout en étant capable de mobiliser les méthodes et les approches de machine learning, bien au-delà de la seule collecte des traces.

[61] Merci à Bilel Benbouzid de nous avoir signalé notamment le rôle joué par la théorie de Vladimir Vapnik qui permet de tester la capacité de généralisation d'un modèle d'apprentissage construit à partir d'un nombre d'exemples finis.

2/ Des traces aux vibrations

Pour rendre plus solides les fondations de ces sciences sociales de troisième génération, il conviendra de donner statut scientifique à ces traces hétérogènes. Or, en préalable, il faut rappeler qu'il est fort probable que toutes ces traces qui pouvaient encore être connectées à des données personnelles ne seront plus accessibles dans les mêmes conditions dans quelques années. Le succès d'Adblock qui bloque les cookies et autres publicités intrusives s'amplifie constamment (200 M téléchargements, 40 % d'installation sur Firefox en 2014). Le cryptage généralisé deviendra une nécessité face à l'incapacité des plates-formes et des services de renseignement à réguler leurs propres activités prédatrices de données personnelles [62]. C'est pourquoi la prise en compte des traces, à la surface même des réseaux, et sans lien avec les données structurées et socio-démographiquement significatives, constitue une base solide de fondation des sciences sociales, au contraire de tous ceux qui continuent de vouloir appliquer leurs modèles de la société et de l'opinion à un univers qui ne leur est accessible qu'en raison d'un laxisme très provisoire. Travailler à la surface de ces traces, sans lien avec les données personnelles, permet aussi de réduire les contradictions éthiques dans lesquelles se trouvent prises les sciences sociales de la société et de l'opinion qui veulent exploiter ces sources.

Nous l'avons souligné, la production des traces est directement dépendante de plates-formes qui génèrent elles-mêmes leurs analyses. Il est cependant nécessaire de fonder les sciences sociales de 3ème génération sur une proposition non captive des utilisations qui sont faites de ces traces, de la même façon que les sondages ne servent pas qu'aux médias ou que les recensements ne servent pas qu'aux états. Aux couples registre/ enquête établi par Alain Desrosières[63], puis audience/ sondages d'opinion, il faut parvenir à ajouter un couple traces/ X, X étant la place qui reste à définir pour la reprise des traces par les sciences sociales. Nous proposons de parler alors de « vibrations ». Le terme permet de filer une métaphore suggestive avec les vibrations des tremblements de terre (aftershock), sachant qu'il est possible de suivre sur le sismographe des répliques qui anticipent et d'autres qui suivent le choc lui-même. Il a l'avantage de focaliser l'attention sur les ondes et moins sur les particules et il fait écho au « buzz » qui obsède les marques et les médias mais qui n'est jamais théorisé.
D'autres concepts pouvaient prétendre à cette place, comme l'attention, l'influence, les « *issues* », l'acteur-réseau, les mèmes, les répliques ou encore les conversations[64]. Cependant, le terme de « vibrations » présente l'avantage d'être familier et polysémique et chacune des acceptions fait sens dans cette science qui traite les traces comme matière première. L'essentiel tient dans le décentrement réalisé vis-à-vis des notions d'acteurs, de stratégies et de représentations, qui ont toutes leur légitimité dans le cadre des autres sciences sociales mais qui ne permettent pas de rendre compte du « pouvoir d'agir » des entités circulantes que sont les vibrations. Nous ne pouvons pas dire *a priori* quelle est la taille ni le statut de ces entités, car ce sont seulement les investigations de corpus de masse qui peuvent nous les faire repérer dès lors que leur réplication émerge des capteurs que nous exploitons, certes à partir des plates-formes mais selon nos objectifs.

---

[62] Bruce Schneier, *Data and Goliath. The Hidden Battles to Collect Your Data and Control Your World*, New-York City, WW Norton and Co, 2015.

[63] Alain Desrosières, *Gouverner par les nombres. L'Argument statistique II*, Paris, Presses de l'École des Mines, 2008. Voir aussi son dernier livre *Prouver et gouverner. Une analyse politique des statistiques publiques*, La Découverte, 2014.

[64] Dominique Boullier, *La télévision telle qu'on la parle. Trois études ethnométhodologiques*, Paris, L'Harmattan, 2004. Dominique Boullier, « La fabrique de l'opinion publique dans les conversations télé », *Réseaux*, 126, 2004, p. 57-87.

Le principe d'une sociologie des vibrations repose sur l'impératif de suivre des éléments (par exemple « je suis Charlie », comme hashtag mais aussi comme icone ou comme citation hors de Twitter), sans pour autant savoir comment ils vont s'agréger pour faire des « tout » à géométrie variable (des groupes très hétérogènes sont pris dans la dynamique de propagation de « je suis Charlie » et aucune frontière ne peut être dessinée). Le parti-pris est donc « élémentariste » mais ne doit surtout pas devenir atomiste car la géométrie variable reste une qualité que nous avons apprise de l'ANT[65]. Toute variation de « je suis charlie » est intéressante, telle que « je ne suis pas charlie » qui relève du même processus, fait d'imitation/opposition, comme le proposait G. Tarde. Comme on le voit, ce n'est pas une substance de l'opinion de supposés individus qui est recherchée mais bien le pouvoir de circulation d'une vibration qui se transforme selon les milieux qu'elle affecte. L'objectif n'est pas non plus de tendre vers la physique sociale, qui cherche des lois supposées transversales à tous ces flux selon des modèles connus en physique des fluides ou en physique corpusculaire[66], car les vibrations sociales gardent leurs particularités et ne s'étudient pas en généralité mais selon les « issues », 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). 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 « issues » que portent et font vivre les vibrations, en ce sens bien différentes des traces « brutes ». L'objet d'étude n'est pas tant l'élément, qui peut avoir des attributs très variés, en étendue et en matérialité, ni seulement les agrégats, ce que l'on tend à faire avec les clusterisations des méthodes de graphes[67], mais bien le processus de circulation et d'agrégation ou de désagrégation, au moment de bifurcation des courbes. Il n'est pas suffisant de repérer les clusters que produit la propagation de « je suis charlie » dans les comptes Facebook ou dans les 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, par exemple. Dans ces courbes, il faut alors plutôt se focaliser sur les moments d'émergence et sur les degrés, comme le préconisait Tarde, et non sur les pics qui fonctionnent, eux, comme des agrégats, comme le fait le memetracker de première génération, ni sur les plateaux comme le faisait Quetelet[68]. L'objet de cette science des vibrations est bien l'agentivité des vibrations qui se propagent et qui finissent par nous prendre, comme l'expérience collective en donne intuitivement l'exemple (avec l'élan mimétique qui « prend » des millions de personnes pour aller manifester le 11 Janvier 2015). Car les individus sont en fait traversés par les idées et les idées nous agissent et non l'inverse comme l'avait bien indiqué G. Tarde[69]. « Les rayons d'imitation d'abord et ensuite des êtres dont on induit l'existence à partir de la variation qu'ils font subir aux flux d'imitation »[70]. Il est alors possible d'étudier les propriétés de ces vibrations pour comparer éventuellement leurs chances de survie ou de contamination rendues possibles par ces différences de propriétés toujours directement liées aux « issues » qu'elles portent avec elles.

[65] Madeleine Akrich, M. Callon et B. Latour, *Sociologie de la traduction. Textes fondateurs*, Paris, Presses des Mines de Paris, 2006.
[66] Alex Pentland, *Social Physics. How good ideas spread. The lessons from a new science*, Penguin Press, 2014
[67] Guilhem Fouetillou, « Le web et le traité constitutionnel européen. Ecologie d'une localité thématique compétitive », *Réseaux*, 147, 2008, p. 229-257.
[68] Gabriel Tarde, 1890, op.cit. p. 173.
[69] Gabriel Tarde, *Monadologie et sociologie*, Paris, Félix Alcan, 1893.
[70] Bruno Latour, « Gabriel Tarde. La société comme possession. La preuve par l'orchestre » in Didier Debaise, *Philosophie des possessions*, Les presses du réel, 2011.

Comme on le voit, l'approche par les vibrations est alors une entrée vers une monadologie (qui se différencie radicalement d'une vision atomiste), ou vers une échologie[71].
La traçabilité n'est cependant pas donnée telle quelle par les plates-formes et nécessite de produire les outils et les méthodes adaptées à une approche des vibrations et non plus seulement des traces. Leskovec et Kleinberg ont fait figure de précurseurs sur ce plan en proposant leur memetracker. Ils sont en effet capables de restituer des flux de tous types, ce qu'ils appellent des *streams* et des *cascades*. Le développement des méthodes de traçabilité des vibrations devra tenir compte de cet existant, en veillant à les tester préalablement sur des *corpora* constitués à cette fin, quitte à en perdre le « réalisme ». Nous avons commencé ce travail en 1987 à partir du suivi des conversations télé transformées sur les lieux de travail pour en faire des « opinions publiques locales »[72]. Nous avons procédé de même pour le suivi des attributs d'une photo dans les bases de données Flickr[73] et des signes transposables dans un corpus de sites web en lien avec une région[74]. Dans le premier cas, ce sont les attributs de la photo (ex : les bras croisés choisis par Roland Barthes comme punctum) qui deviennent des attracteurs de tags et qui connectent ainsi des comptes ou des photos qui n'auraient jamais été reliés a priori selon ces critères. Mais nous nous sommes arrêtés au principe de ce travail sans pouvoir le conduire empiriquement à une échelle suffisante. Dans le second cas, la propagation du drapeau breton sur le web devient un indicateur de connexion que l'on peut comparer avec d'autres entités constitutives de l'imagerie régionale qui, elles, ne parviennent pas à se propager. A partir d'un travail manuel sur près de 600 sites réalisé par M. Le Béchec, il fut possible d'esquisser ce que serait l'analyse des vibrations générées par ce drapeau et de constater que ses propriétés sémiotiques n'étaient pas étrangères à sa capacité de circulation. Tags ou icônes sont ainsi des vibrations que l'on peut suivre, quand bien même ils n'ont pas le caractère explicite de verbatims ou d'expressions comme dans le memetracker ni leur caractère massif. Potentiellement, toutes les traces que nous avons identifiées (telles que les likes, les tweets, les recommandations, etc.) sont susceptibles de faire l'objet de ces suivis : ils nécessitent cependant des outils de traçabilité spécifiques, qui existent en grande partie pour Twitter seulement mais à la condition de faire un réexamen détaillé de ces outils pour vérifier qu'ils répondent au cahier des charges d'une traçabilité des vibrations (et non seulement des traces pour elles-mêmes ou pour la réactivité des marques).

Ce cahier des charges d'une nouvelle génération de sciences sociales parait constituer un objectif inatteignable tant les enjeux méthodologiques et politiques (quels liens avec les plates-formes ?) semblent énormes. Pourtant, ce moment historique qui perturbe le statu quo des façons de faire des sciences sociales n'est pas inédit. Si nous avons dénommé cette approche « troisième génération », c'est pour mieux mettre en valeur sa parenté et ses différences avec deux autres générations que nous avons présentées dans le tableau initial. Désormais il n'est plus possible de contester l'idée de « société », qui existe indépendamment de l'histoire conceptuelle et des dispositifs qui l'ont stabilisée. Et pourtant, comme l'a montré Alain Desrosières[75], il a fallu de longues années de construction des statistiques pour faire

[71] Terme utilisé par Yves Citton après Gilles Deleuze (Yves Citton, *Pour une écologie de l'attention*, Paris, Seuil, 2014). Y. Citton mobilise aussi le concept de vibrations pour l'étude littéraire de l'influence de Spinoza, voir Yves Citton, « Le réseau comme résonance : présence ambiguë du spinozisme dans l'espace intellectuel des Lumières », in Wladimir Berelowitch & Michel Porret, *Réseaux de l'esprit en Europe, des Lumières au XIXe siècle*, Droz, Genève, 2009, p. 229-249
[72] D. Boullier, *La télévision telle qu'on la parle*, op cit., 2004.
[73] Dominique Boullier et Maxime Crépel, « Biographie d'une photo numérique et pouvoir des tags : classer/circuler », *Revue d'Anthropologie des Connaissances*, 7 (4), 2013, p. 785-813.
[74] Mariannig Le Béchec et Dominique Boullier, « Communautés imaginées et signes transposables sur un "web territorial" », *Etudes de communication*, 42, 2014, p.113-125.
[75] Alain Desrosières, 1993, op.cit.

exister quantitativement cette société et la rendre indépendante des concepts de Durkheim. De même, l'opinion publique parait désormais posséder une réalité indiscutable. Pourtant, ce sont les sondages d'opinion qui lui donnèrent cette force d'évidence avec Gallup[76] à partir de 1936. Certes, ils restent contestés mais l'idée même d'opinion publique a fini par devenir « évidente » et c'est en cela qu'ils ont gagné la partie. Comme le dit Bruno Latour, un objet tient parce qu'il a été bien construit et à ce moment il nous dépasse[77]. Les sciences sociales de troisième génération doivent prendre leur place à côté des autres sciences sociales de la société et de l'opinion et non prendre leur place. Car c'est une autre strate du social qui affleure désormais grâce à ces dispositifs de traçabilité à haute fréquence. Mais pour cela, il faudra tirer les leçons des méthodes de construction de conventions qui ont permis aux autres générations de durer, bien au-delà de la prolifération de recettes non contrôlées que l'on observe actuellement. L'insatisfaction vis-à-vis de la qualité des résultats obtenus en pratiquant ce « positivisme algorithmique » que nous avons évoqué est aussi présente chez les marques ou les agences soucieuses de la qualité de leur travail. Ce ressort constitue sans doute la chance des sciences sociales de trouver un accord pour construire une convention avec toutes les parties prenantes, pour que des principes issus des exigences scientifiques soient en même temps producteurs de qualité des résultats d'un point de vue opérationnel.

Conclusion

La construction d'une offre de sciences sociales de troisième génération n'est pas garantie. Nous avons voulu en proposer ici les conditions de félicité dans la lignée de la théorie de l'acteur-réseau qui en a posé les prémisses et de Tarde qui en avait annoncé les principes. Mais pour l'instant, la tendance à la fin de la théorie et à l'occupation du terrain par les plates-formes du web (GAFAT) qui elles-mêmes produisent, calculent, et publient sur ces traces, reste dominante, et cela pour des visées commerciales avant tout, puisque les marques sont les grands demandeurs de ces approches. Rappelons que nous reconnaissons l'intérêt pour les marques d'apprendre à réagir en utilisant ces métriques. De même que nous reconnaissons l'intérêt pour les sciences sociales de la société et de l'opinion de continuer à développer leurs approches en utilisant ces sources. En ce sens, nous plaidons pour faire coexister ces approches, pour apprendre à changer de point de vue de l'une à l'autre[78] et pour admettre les conditions de possibilité de chaque génération, appuyées sur les Etats, les médias ou les marques. Chaque étude spécifique d'une question issue de l'expérience ordinaire ou posée par ces prescripteurs doit conduire à combiner les trois générations sans pour autant prétendre à la totalisation. A la condition que la recherche dispose d'un cadre spécifique pour ces traces qui envahissent notre monde. A chaque longueur d'onde sociale, ses méthodes et ses limites de validité. Cela instituera un principe de précaution salutaire dans l'usage de ces traces[79]. C'est

[76] George Gallup, *Public Opinion in a Democracy*, Herbert L. Baker Foundation, 1939, Stafford Little lectures.
[77] Bruno Latour, *Petite réflexion sur le culte moderne des dieux faitiches*, Paris, Les empêcheurs de penser en rond, 1996
[78] Cette approche conduit à prolonger le travail de rassemblement du social que Bruno Latour a engagé, dans une veine ici plus diplomatique. Les sciences sociales qui produisent de la société de niveau 2 conservent en effet toute leur place puisque précisément, comme toute la sociologie des sciences nous l'a appris, elles sont parvenues à faire tenir leurs principes comme conventions naturalisées. Bruno Latour, *Reassembling the Social - An Introduction to Actor-Network-Theory*, Oxford, Oxford University Press, 2005. Traduction française : Bruno Latour, *Changer la société. Refaire de la sociologie*, Paris, La Découverte, 2006.
[79] Nous rejoignons ici la conclusion de Claire Lemercier, 2015, op. cit. « Discuter et enrichir les modèles causaux des autres sciences sociales, en particulier en pensant *ensemble* les différentes temporalités : le

en cela que les sciences sociales de troisième génération peuvent à la fois aider à rendre compte de phénomènes inédits et faire préciser pour chaque génération son domaine de validité pour éviter toute esprit « cause-finalier » que Tarde exécrait. Notre intention est seulement de contribuer à poser les bases d'une convention, d'un investissement de forme[80], permettant de faire émerger une théorie sociale et un objet, les vibrations, qui ne réduisent pas le numérique aux « méthodes numériques » ni aux « humanités numériques ». Il existe une nouvelle matière première qui mérite un examen pour elle-même et qui produit une troisième strate du social, mesurable selon d'autres principes, et non réductible à la société ou à l'opinion. La société a fini par exister, l'opinion a fini par exister, les vibrations doivent finir par exister au même titre.

programme reste d'actualité. » Mais nous insistons sur des générations constituées historiquement et non sur des disciplines, sur les objets distincts ainsi traités et sur l'impossibilité de penser « ensemble » ces temporalités mais à tour de rôle.

[80] Laurent Thévenot, "Les investissements de forme" in Thévenot L. (ed), *Conventions économiques*, Paris, CEE-PUF, 1986, pp. 21-71

Glossaire

Big Data

Le terme fait référence à la fois au volume de données disponibles pour des calculs de tous types, mais aussi à leur variété puisqu'il est désormais possible de combiner des données enregistrées par des services officiels classiques (comme les demandes d'emploi à Pôle Emploi), des traces issues d'activité sur les réseaux (des requêtes Google pour réaliser son CV), et des verbatims récupérés sur des forums ou sites de presse par exemple. Leur mise à jour constante (vélocité) grâce aux réseaux confère à ces données un statut différent de celui des bases de données dont les sciences sociales sont familières. Mais le terme fait aussi référence aux méthodes de calcul qui exploite ces sources de données et qui relèvent du machine learning et de l'inférence à partir de corrélations testées en masse.

High Frequency Trading

Le High Frequency Trading ou transactions à haute fréquence fait partie des techniques financières de trading automatique, c'est-à-dire directement gérées par les ordinateurs et leurs algorithmes. Il permet de jouer sur les écarts entre ordres de vente et d'achat et de tirer profit de ces écarts à la microseconde près. Il cherche à influencer le carnet d'ordres des autres investisseurs en générant artificiellement des tendances à la hausse et à la baisse par exemple en inondant le marché d'ordres longs à calculer pour les concurrents et finalement non réalisés. Cette technique éminemment spéculative nécessite les machines et les réseaux les plus performants.

Verbatims

Les transcriptions des entretiens bien connus des sciences sociales sont ici étendus à des extraits d'expressions récoltées dans des environnements divers (forums, posts de réseaux sociaux, avis de consommateurs, etc.) sous forme écrite non préformatée mais calculable par des algorithmes de traitement automatique du langage naturel (TALN). Ils constituent désormais

Timestamp ou horodatage

Chaque opération réalisée sur une machine informatique et sur les réseaux se voit attachée une métadonnée indiquant sa date et son heure exacte de réalisation et collectée dans un journal d'événements. Indépendamment des contenus de cette opération, cette seule métadonnée est calculable et permet de réaliser des calculs variés.

Timeline ou ligne de temps

La timeline est un mode de visualisation graphique de données à caractère chronologique, sur une même ligne de temps, qui peut devenir une frise chronologique comportant des aspects interactifs pour accéder directement aux documents associés aux événements.

Humanités numériques

Le label « humanités numériques » regroupe toutes les combinaisons possibles entre sciences humaines et sociales et informatique. Elles comportent la dimension de numérisation de sources, la mobilisation de techniques de captation et d'analyse numériques adaptées aux corpus intéressant les SHS, la constitution de corpus partagés et permettant des expérimentations, jusqu'à l'exploitation des nouvelles sources de données produites par les réseaux numériques et la mobilisation de techniques de calcul et de visualisation inédites.

Amplification
Le concept a été utilisé par Elisabeth Eiseinstein dans son analyse historique de l'imprimé pour montrer que toutes les tendances déjà présentes dans les sociétés de l'époque ont été amplifiées à la fois par l'imprimerie, avant que puisse être constaté longtemps après que certaines de ces tendances avaient bénéficié plus nettement de cette nouvelle technologie. Ce concept permet de relativiser la notion de révolution appliquée trop souvent au numérique sans que l'on sache quelles tendances seront finalement gagnantes dans la longue durée. Il permet de rendre compte aussi de la prolifération des innovations et des usages liés à ces innovations technologiques qui provoque une forme de désorientation mais aussi rend possible un débat politique sur les choix techniques à favoriser.

Mèmes
Le concept de même a été proposé par R Dawkins dans « The selfish gene » pour étendre la théorie évolutionniste depuis la génétique vers des entités culturelles élémentaires que sont les mèmes, qui présentent la particularité de se propager par imitation et par dérivation. L'analyse de tous les processus culturels les plus fondamentaux (le langage, le soi, le cerveau) sous cet angle a été tentée (Blackmore) avant d'être relativement délaissée. Les phénomènes de viralité rencontrés sur le web ont remis cette notion à l'ordre du jour au point d'en faire un mode de production reconnu et suivi, à partir d'images reproduites et modifiées, de termes à haut pouvoir de contagion, qui participent à la constitution de ce phénomène appelé « buzz », ici analysé comme vibrations.

Machine learning
Le machine learning est une nouvelle version de l'intelligence artificielle qui ne s'appuie plus comme au XXeme siècle sur une catégorisation exhaustive du monde en ontologies a priori pour effectuer les calculs (ce qui n'est possible que dans certains univers très standardisés comme l'industrie aéronautique). Désormais, les machines peuvent apprendre des données qui leur sont fournies en flux permanent et en quantité suffisante (Big Data) pour tester les corrélations entre elles. Le machine learning mobilise des modèles d'apprentissage multiples et non plus des modèles du monde, il sélectionne les algorithmes pertinents au sein de bibliothèques d'algorithmes en les testant selon les types de données et procède par apprentissage supervisé, incluant les retours et validations des experts du domaine. Cette flexibilité extrême le rend adaptable à tout contexte, mais suppose des capacités de calcul et de stockage énormes mais désormais disponibles à coûts raisonnables pour les grandes entreprises.

Algorithme
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 et non ambigus. C'est à cette condition qu'il peut être exécuté par des machines informatiques grâce à leur écriture sous forme de code.

Score d'autorité
Lorsqu'on collecte les liens (arcs) entre sites (nœuds) dans un réseau, numérique ou non, il est possible de calculer un score d'autorité pour chaque nœud à partir du nombre de sites qui pointent vers lui, des liens entrants (in-degree). Le score d'autorité est un des éléments de la topologie du web exploités par Google pour produire son ranking des pages.

Score de hub

Lorsqu'on collecte les liens (arcs) entre sites (nœuds) dans un réseau, numérique ou non, il est possible de calculer un score de hub pour chaque nœud à partir du nombre de sites vers lesquels il pointe, des liens sortants (out-degree). Le score de hub est un des éléments de la topologie du web exploités par Google pour produire son ranking des pages.