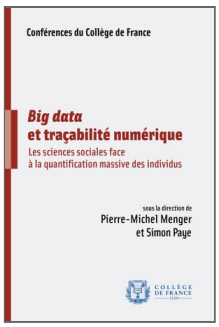


Pierre-Michel Menger et Simon Paye (dir.)

Big data et traçabilité numérique
Les sciences sociales face à la quantification massive des individus

Collège de France

# Pour des sciences sociales de troisième génération (SS3G)

## Des traces numériques aux réplications

Dominique Boullier

Éditeur : Collège de France
Lieu d'édition : Paris
Année d'édition : 2017
Date de mise en ligne : 23 octobre 2017
Collection : Conférences
ISBN électronique : 9782722604674

**Édition imprimée**
Date de publication : 24 octobre 2017



**Référence électronique**
BOULLIER, Dominique. *Pour des sciences sociales de troisième génération (SS3G) : Des traces numériques aux réplications* In : *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus* [en ligne]. Paris : Collège de France, 2017 (généré le 24 octobre 2017). Disponible sur Internet : <http://books.openedition.org/cdf/5011>. ISBN : 9782722604674.

# Pour des sciences sociales de troisième génération (SS3G)

## Des traces numériques aux réplications

**Dominique Boullier**
Sociologue, professeur à l'École polytechnique fédérale de Lausanne 3

### 1. L'âge du numérique

Tim Berners-Lee est sans aucun doute un grand bienfaiteur de l'humanité pour avoir accepté de livrer le code HTML à tous, en le laissant ouvert et en lui permettant ainsi de provoquer cette interconnexion généralisée des contenus. Il en donne d'ailleurs une vision historique intéressante lorsqu'en 2008, il propose ce récit du passage du III au WWW, puis au GGG. « III » vaut pour « Internet International Infrastructure » et fut mis en place à partir de 1974. Cette capacité à contrer la supposée toute-puissance des centraux des opérateurs de télécommunications au profit d'un réseau distribué, dans lequel toute machine peut devenir un serveur, doit être soulignée comme un véritable coup de force en faveur d'un réseau neutre qui a produit des effets proliférants remarquables avant d'être remis en question par de nouvelles formes de centralisation et de hiérarchie. Les machines étaient reliées entre elles grâce au protocole IP. L'idée de Tim Berners-Lee et de Robert Cailliau en 1990 fut d'étendre ce principe de réseau distribué aux documents eux-mêmes, ce qui donna ce protocole HTML, porteur de l'hypertexte, depuis longtemps imaginé mais non implémenté, qui déboucha sur le World Wide Web. Les documents étaient reliés entre eux et tous accessibles par un même principe, leur URL (*Uniform Resource Locator*), qui correspond à leur localisation unique dans le réseau, indépendamment des machines et des types de réseaux.

Rendre ces contenus accessibles à tous par un tel système de balisage contestait de fait l'autorité des bibliothèques ou de tout autre centre de ressources ou base de données, comme Internet contestait l'autorité des opérateurs de télécommunications. Il fallut cependant mettre en place des moteurs de recherche, et non plus des annuaires qui décalquaient ces autorités de classement, pour mesurer toute la puissance potentielle du WWW. Berners-Lee conteste cependant la domination de ce modèle car il prétend que « Le Web ne relie pas seulement les machines, il relie des personnes[1]. » Pour lui, l'ère actuelle est celle du GGG, Global Giant Graph ou « graphe géant global », celle de la connexion des personnes, non seulement à travers les réseaux sociaux, mais plus généralement parce que toute adresse IP (du III) ou tout document (du WWW) renvoie toujours à des personnes (dans le GGG), toutes mises en réseau. Les informaticiens les plus géniaux sont capables d'assumer leur rôle de connecteurs sociaux et tentent d'en faire la théorie, sans s'em-

1. « *The Web does not just connect machines, it connects people.* » Discours à la Knight Foundation, 14 septembre 2008, https://webfoundation.org/about/community/knight-2008-tbl-speech.

barrasser des concepts de la tradition sociologique, on s'en doute. Il est aisé d'observer comment ces services de réseaux sociaux ont préempté des notions telles que « communautés » ou « amis » selon leurs propres critères, et avec une telle puissance de frappe que leur propre interprétation tend à s'imposer comme évidente. Nous voudrions montrer à quel point, dans cette approche, une tradition sociologique que nous qualifierons de « première génération » perdure en traitant le numérique comme un terrain supplémentaire de démonstration de la force de la « société ». Nous prétendons ensuite qu'un autre modèle émerge dans ce GGG, fondé sur les traces, qui permet de relier humains et non-humains, comptes de réseaux sociaux aussi bien que capteurs sur des objets de plus en plus nombreux et qui agissent pour leur compte. La sociologie ne pourra saisir ce phénomène émergent qu'à la condition d'accepter d'inventer un tout autre cadre, celui qui permet de passer de ces traces aux réplications. Les réplications permettent de penser la portée sociologique de ces traces numériques et d'aborder ainsi un phénomène jusqu'ici impossible à suivre à la trace mais rendu accessible par les *big data*[2] (données et traces calculables en masse).

### 1.1. Ni personnes ni identités, les traces sont la matière première

Reprenons déjà cette approche de Berners-Lee pour montrer en quoi elle est trop sociologique et trop peu précise quant aux propriétés de ce qui est assemblé dans ce graphe géant global. Facebook a certes réussi un tour de force incroyable en rendant quasi obligatoire (ou tout au moins normal du point des acteurs eux-mêmes) de déclarer son identité véritable, c'est-à-dire celle fournie par l'état civil, son nom et son prénom. Pourtant tout le Web avait été marqué avant cela par l'anonymat, considéré comme un avantage et une forme de préservation de son droit d'expression libre, avec toutes les dérives que d'autres critiquaient pour les mêmes raisons. Or, les identités de Facebook ne renvoient à des personnes identifiables par l'état civil que grâce à cet effet de normalisation car, techniquement, rien ne permet de garantir quelque lien que ce soit (puisque seule une adresse mail certifie cette identité lorsqu'il y a vérification). Depuis, les « fausses » identités se sont d'ailleurs multipliées malgré le souhait de Facebook d'organiser « le grand déménagement numérique de l'état civil » à son profit, puisqu'on peut désormais se connecter avec son seul compte Facebook comme garant de son identité. Ce ne sont pas des personnes qui sont connectées sur ce réseau social mais des traces d'activité d'une entité qui peut prendre éventuellement les formats de l'état civil. Il convient de rester au plus près de ces propriétés techniques, de ces traces, pour comprendre ce qui se joue sur les réseaux numériques en général et pour éviter toute image d'un Internet, d'un Web ou d'une application de réseau social comme « papier carbone » d'une société ou d'entités sociales aisément identifiables, ce que toutes les méthodes de représentation de la société ont toujours cherché à faire. Il est plus aisé de le voir dans le cas de Google, autre plate-forme qui a formaté désormais toutes nos relations avec le monde des documents, celui du WWW. Dans le cas du moteur de recherche, aucun intérêt pour les personnes n'est nécessaire ni même pour les sites, supposés porteurs de contenus renvoyant à des autorités, à des communautés ou à d'autres entités sociales couramment analysées dans

2. Des versions remaniées de ce chapitre ont été publiées dans la revue *Socio* (Boullier, 2015a) et dans la *Revue française de science politique* (Boullier, 2015b).

les sciences sociales. Les scores qui permettent de classer les sites reposent sur une topologie qui ne traite jamais de leurs contenus en tant que tels, et où les liens entrants et les liens sortants produisent un rang d'autorité ou de *hub*, mais au sens de la topologie des réseaux (Kleinberg *et al.*, 1998) et non d'un statut social. De même, ce classement ou *ranking* topologique, lié à la notoriété (*hub* ou autorité), est affecté notamment par l'audience d'un site donné, selon un nombre de clics, c'est-à-dire à un niveau de traces très bas, sans référence aux propriétés des personnes identifiables par leur état civil. Dominique Cardon a proposé de distinguer les métriques du Web portant sur la vue (le fait d'avoir vu une page), le lien (hypertexte), le *like* et la trace (Cardon, 2013).

Précisons ici d'emblée ce que nous entendons par « traces », traces qui intègrent tous les éléments proposés par Cardon. Elles se distinguent en effet des données que l'on peut récupérer en masse sur des fichiers clients ou encore à partir d'actes administratifs. 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 que la sociologie ou le marketing sont plus habitués à mobiliser (attributs socio-démographiques). Les traces peuvent aller de signaux (bruts) à des verbatims non structurés, elles peuvent être des traces (telles que les liens, les clics, les *likes*) qui sont exploitées en bases de données par les opérateurs mais qui sont aussi captées indépendamment de cela à travers les API[3] et qui ne relèvent pas alors de bases de données relationnelles. Toutes ces traces et les calculs qui y sont appliqués constituent une partie du phénomène Big Data. Le traitement de ces données/traces mobilise des applications de structuration des données massives ou Big Data Architecture Frameworks (BDAF) massivement parallèles. Mais les traces peuvent aussi comporter tous les flux de données entre machines et entre objets qui seront bientôt, grâce au protocole IP version 6, dotés d'adresses IP ($3{,}4 \times 10^{38}$ adresses disponibles) qui leur donnent un statut équivalent aux autres traces, apparemment plus humaines. Les traces sont donc des données détachées et détachables de leurs contextes de production et de calcul : c'est en cela qu'elles ont un rapport avec les *big data* car elles ne sont pas nécessairement préformatées pour un calcul précis ni dépendantes de l'agrégation que l'on peut appliquer ensuite. Il est aisé de dire que malgré tout, « derrière » les sites ou « derrière » les clics, il y a bien des humains, mais cela n'enlève rien au fait que les algorithmes, eux, ne s'intéressent pas à cette propriété et que, de plus, aucune certitude ne peut être apportée sur ce plan. 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 ne sont pas créées avec d'autres attributs.

Notons cependant que les métadonnées attachées à ces traces comportent automatiquement un *timestamp* (horodatage), qui permet de produire une *timeline* (un historique) qui semble un premier attachement quasi évident. Elles comportent de plus en plus un *tag* de géolocalisation (effectif ou déduit de l'adresse IP, avec toutes les approximations que cela suppose), ce qui sert très rapidement à produire des cartes, dont l'extension est permanente dans le domaine de traitement de ces traces et à portée de tout internaute grâce à l'omnipré-

3. Les *application programming interfaces* permettent de se connecter aux bases de données des plates-formes et d'en utiliser certains éléments pour réaliser des calculs et des applications alors que le logiciel reste propriétaire et non accessible.

sence des Google Maps ou de OSM (Open Street Map), deux ressources stratégiques pour s'orienter dans la prolifération des traces. Cela indique bien que l'opération de détachement reste limitée et que ces deux ressources, le temps et l'espace, suffisent à faire émerger des corrélations interprétables de façon stéréotypée voire caricaturale (les requêtes de recettes de cuisine avant Thanksgiving dans les différents États américains) ou ouvertes à tous les tests.

Pour Amazon ou Apple (puisque le Web n'est plus distribué mais bien accaparé par ces quatre plates-formes GAFA qui concentrent une part importante du trafic, par exemple 6,4 % pour Google), ce ne sont pas non plus des personnes qui sont mises en relation mais avant tout des goûts (livres ou musique à l'origine), exprimés par des traces d'achat, de préférences, qui peuvent être traitées en masse pour produire des patterns, des profils, indépendamment des informations personnelles. Berners-Lee a raison de considérer le passage du WWW au GGG comme un moment décisif, à condition de ne pas rabattre le graphe sur des supposées personnes, mais seulement sur des attributs, plus ou moins connectés. Certes, il convient de ne pas oublier que toutes ces plates-formes sans exception sont aussi très friandes de données de type état civil, numéros de téléphones et autres ressources très intéressantes pour les annonceurs à qui elles les revendent. Les scandales sur le non-respect de ces données personnelles sont devenus une constante de la vie du Web. Mais les algorithmes qui ont fait leur fortune reposent avant tout sur les traces laissées – volontairement ou non – par les internautes, des traces de bas niveau (un clic) qui, une fois conservées, comparées (matchées) et modélisées, donnent déjà beaucoup d'informations sur les tendances d'un marché particulier, sur les publics d'un site, etc. Et les méthodes de marketing qui en découlent reposent plus largement sur l'adressage de masse de publicités ou de mails à des adresses IP qui ont cliqué sur un article (*retargeting*), que sur des mises en relation sophistiquées avec les autres attributs des supposées personnes attachées à ces adresses ou à ces clics (*profiling*). L'investissement est massif dans ces secteurs pour parvenir à faire le lien à partir des nouveaux avatars de l'intelligence artificielle (*machine learning*), mais le chemin est encore long avant d'en produire des résultats et des stratégies pertinentes.

### 1.2. Les traces produites par des plates-formes

La matière première de ces plates-formes, ce sont bien des traces numériques, que l'on peut étendre à tous les commentaires, les *likes*, ou autres étoiles de recommandation qui font l'activité quotidienne des internautes. Dès lors, les sciences sociales font face à une alternative : soit elles se retrouvent cantonnées à un autre monde en relativisant l'intérêt de ce type de traces et en privilégiant les données, soit elles décident de chevaucher le tigre et de prendre ces traces comme matière première à leur tour. Elles doivent alors accepter de dépendre des plates-formes qui produisent ces traces, sans pouvoir peser d'un quelconque poids sur leur formatage, voire en dépendant totalement des conditions de fourniture de ces données. Les chercheurs qui veulent faire des requêtes sur les données de Twitter (le plus ouvert), de Facebook ou de Google, savent bien par exemple qu'ils sont limités en volume et qu'ils doivent le plus souvent se contenter des outils d'exploration fournis par ces plates-formes elles-mêmes (Google Scholar, Google Trends, Facebook Social Graph, etc.). Les limites de la qualité des traces sont observables sur toutes les plates-formes, mais ces limites peuvent être intrinsèques lorsqu'elles ne répondent pas au critère

de traçabilité que nous considérons comme décisif pour les exploiter, ou extrinsèques lorsqu'on critique leur absence de relation fiable avec le monde « réel », celui qui fait la « société ». C'est cette dernière posture que l'on trouve chez Boyd et Crawford à propos de Twitter :

> Certains utilisateurs ont des comptes multiples. Certains comptes sont utilisés par de multiples utilisateurs. Certains internautes ne créent jamais de compte, et accèdent simplement à Twitter *via* le Web. Certains comptes sont des robots qui produisent des contenus générés automatiquement sans faire intervenir une personne. De plus, la notion de compte « actif » est problématique. Alors que certains utilisateurs postent du contenu régulièrement sur Twitter, d'autres contribuent en tant qu'« observateurs ». Twitter Inc. a révélé que 40 % d'utilisateurs actifs s'inscrivent uniquement pour « observer[4] ». (Boyd et Crawford, 2011)

D'autres travaux (Driscoll et Walker, 2014) ont testé les données produites à partir de différentes méthodes d'accès offertes par Twitter par exemple et ont montré que l'API Search, l'API Streaming et le Gnip Power Track (service payant) fournissent des résultats très différents, la dernière méthode récoltant un bien plus grand nombre de tweets, mais pas de façon uniforme selon les requêtes ! C'est dire que les traces collectées sont entièrement dépendantes des dispositifs de collecte, ce qui ne saurait étonner mais qu'on a tendance à oublier lorsqu'il s'agit d'autres méthodes plus anciennes qui sont devenues conventionnelles. Verra-t-on ainsi l'émergence de « Google sciences », de « Facebook sciences » ou de « Twitter sciences » tant la dépendance serait forte à ces formats ?

### 1.3. L'emprise des marques sur les traces

Pourtant si ces traces sont devenues aussi recherchées, ce n'est pas d'abord à cause de leur intérêt pour les sciences sociales évidemment mais parce qu'elles sont une des ressources clés pour les marques pour suivre les effets de leurs propres actions sur leur public. La réputation, la notoriété, ne se traduisent plus seulement dans des mesures d'audience qui seraient une importation simpliste des mesures longuement construites pour les médias de masse (*mass media*). Sur les réseaux, il faut mesurer à la fois une forme d'audience (le *reach*), des activités les plus élémentaires de 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 ; ce sont elles qui alimentent les recettes de toutes ces plates-formes et, par-là, de tout le Web. Les marques ont envahi la sphère médiatique depuis trente ans, dès lors qu'il a fallu investir fortement dans la communication et le marketing pour contrer une baisse de la consommation provoquée par une pression constante sur le pouvoir d'achat. Le triomphe du marketing n'a pas d'autre origine que cette nécessité de maintenir des parts de marché dans des environnements de plus en plus concurrentiels, puisqu'à la fois

4. « *Some users have multiple accounts. Some accounts are used by multiple people. Some people never establish an account, and simply access Twitter via the Web. Some accounts are "bots" that produce automated content without involving a person. Furthermore, the notion of an 'active' account is problematic. While some users post content frequently through Twitter, others participate as "listeners". Twitter Inc. has revealed that 40 percent of active users sign in just to listen.* »

les ressources des ménages se réduisent et les exigences de marge de la part des investisseurs augmentent. Ces propriétés bien connues de l'économie financière la connectent directement avec le monde des médias puis des réseaux numériques pour en faire une économie d'opinion (Orléan, 2011). L'inquiétude permanente des professionnels du marketing et de la communication devant les flux incontrôlables des avis et des réactions agrégées sur le Web leur impose une forme de thérapie, agissant sans doute largement comme un placebo, mais efficace pour cette raison même, celle du suivi des réputations, des opinions, de ces entités encore incertaines qui prolifèrent et se propagent sur le Web. Les corrélations avec des taux de conversion, c'est-à-dire des actes d'achat effectifs, sont beaucoup plus incertaines, techniquement complexes et rarement exploitées, car l'économie de la réputation des marques vise tout autant les investisseurs que les clients des produits ou services. Les outils d'*opinion mining* et de *sentiment analysis* (analyse des sentiments) que nous avons examinés en détail (Boullier et Lohard, 2012) constituent ainsi la réponse à cette angoisse du marketeur après le lancement de produit. Cependant, l'extension de ce domaine de la marque atteint toutes les activités, qu'elles soient commerciales, culturelles, politiques, institutionnelles voire interindividuelles lorsque chacun doit mesurer son excellence à l'aide de *rankings*, comme les chercheurs sont poussés à le faire. Dès lors, ce sont les méthodes des marques qui prennent le dessus et imposent leur loi et leur rythme. 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 conduits à considérer que nous étions entrés dans l'ère du High Frequency Politics (Boullier, 2013) à l'image du High Frequency Trading de la finance spéculative.

### 1.4. Que viennent faire les sciences sociales dans cette galère ?

Nous avons ainsi dressé un tableau qui mérite d'être systématisé. Le numérique en réseaux génère :

- des traces ;
- assemblées et formatées par des plates-formes ;
- pour des marques ;
- en vue d'une réactivité ;
- pour produire des *rankings* ou des *patterns*.

Que peuvent bien faire les sciences sociales de telles ressources ? Voilà l'enjeu que nous voulons mettre en avant, sachant que le risque est de déléguer l'exploitation de ces traces aux plates-formes elles-mêmes pour qu'elles deviennent les nouveaux médiateurs de la réflexivité de nos sociétés sur elles-mêmes.

Cette situation n'est pas nouvelle, la mutation technique semble plus directement en cause dans cette chaîne de médiations qui s'alignent, mais deux autres moments clés de l'existence des sciences sociales et en particulier de la sociologie doivent être mis en parallèle selon la même méthode pour comprendre la portée des changements en cours.

## 2. La construction de l'« opinion[5] »

La situation contemporaine n'est sans doute pas si éloignée d'un moment-clé dans l'histoire des sciences sociales qui nous aiderait à comprendre ce qui se passe. Si l'on donnait à l'époque actuelle des traces numériques le libellé de « 3G », pour troisième génération, il faudrait alors donner à l'émergence de l'opinion à la fin des années 1930 le libellé de 2G. En 1936 en effet, George Gallup parvint à prédire l'élection de Roosevelt face à Alfred Landon avec une étude sur 50 000 personnes. Elmo Roper et Archibald Crossley avaient fait de même au même moment. Non seulement il impressionna les médias et les décideurs, mais il disqualifia radicalement les méthodes anciennes (*straw polls*), celles du *Literary Digest* fondées sur les réponses de 2 millions de personnes, en prédisant même leurs propres résultats erronés. Ce qu'il fondait ainsi dans ce geste spectaculaire, c'était la fiabilité du sondage et des méthodes d'enquête par échantillonnage, le *sampling*, qui certes perdait l'*exhaustivité* des enquêtes sur une population entière, mais parvenait à des résultats corrects à condition de respecter des conditions de *représentativité*. Il échouera cependant en 1948 à prédire la victoire de Truman, dont les électeurs se décidèrent dans les dix derniers jours. Les méthodes ainsi appliquées à la vie politique et à une épreuve grandeur nature aussi importante qu'une élection présidentielle avaient été testées auparavant sur les études de lectorat pour lesquelles Gallup avait rendu opérationnel l'échantillonnage stratifié. L'opération de légitimation de l'échantillonnage réussit en général grâce aux performances de Gallup, entièrement dédiées à d'autres mondes sociaux, ceux de l-« opinion publique », et non plus de la « société », qui restait la référence des statisticiens de l'État fédéral et de ses bureaux, ceux-ci travaillant aussi à produire des règles d'échantillonnage aléatoire (Didier, 2009). C'est bien dans le contexte des médias de masse que leur importance fut reconnue. Avec David Ogilvy en effet, Gallup étudia les audiences des films puis, chez Young & Rubicam, avec Crossley, les audiences de la radio à partir d'entretiens téléphoniques avant même de proposer ces sondages électoraux. Le nom de Gallup doit être de ce point de vue associé à celui de Paul Lazarsfeld, qui, dans la même période, en 1936, lançait un Radio Research Program, fondé sur ses travaux d'étude d'audience de la radio commencés en 1930. Avec Merton, ils lancèrent les méthodes de *focus groups* ou « groupes de discussion » dès 1941, et l'étude de Decatur en 1945 fournit les données pour l'analyse de *Personal Influence*, publié en 1955 (Katz et Lazarsfeld, 1955), qui établit le cadre d'analyse du *two-step flow* (communication à double étage) dans lequel les médias de masse jouent un rôle, mais à travers les médiations des relations d'influence de divers types.

### 2.1. Les médiations qui font exister l'opinion publique sont constituées

Le lien entre les médias de masse et la vie politique est ainsi constitutif des nouvelles méthodes statistiques d'échantillonnage stratifié (certes fondées sur des quotas et non aléatoires). Ainsi que le note Alain Desrosières (1993), la condition de *prédictibilité* d'une élection nationale dépendait en fait de la constitution d'un espace public médiatique commun à l'échelle des

5. Les travaux de Loïc Blondiaux (1998) et de Joëlle Zask (2000) développent cette histoire largement en français.

États-Unis, et seule la radio pouvait le faire de façon à rendre comparable l'état de connaissances des électeurs à propos des différents candidats. Une mutation médiatique considérable, les médias de masse (la radio à l'époque), a donc constitué les conditions d'émergence et de validation d'une technique d'enquête, qui ouvre ainsi toute une nouvelle époque, notamment pour la science politique. Plus encore, c'est l'« opinion publique » elle-même qui prend une existence mesurable, par ces méthodes d'échantillonnage dont la puissance performative dépassera largement la phase expérimentale.

### 2.2. Des marchés et des publics nationaux : les échelles des médias

Le maillon manquant dans toute notre description reste en effet le levier d'intéressement financier à de tels investissements pour connaître un public. Les agences de communication comme les instituts de sondage ne peuvent en effet vivre de leurs seules activités électorales quand bien même elles leur apportent une grande visibilité et une grande notoriété. Leur cible est au départ constituée par les médias de masse, disions-nous, pour une raison essentielle : la mesure d'audience devient la clé de répartition des espaces publicitaires, et cela dès l'origine avec la radio puis avec la télévision (en 1941 sont diffusées les premières publicités à la télévision américaine pour les montres Bulova pendant un match de baseball). Mais ces mesures permettent aussi de suivre les effets de ces campagnes publicitaires sur les esprits des consommateurs, donnant un essor sans précédent au marketing qui pilote des stratégies de communication de plus en plus sophistiquées à l'échelle d'un pays (Cochoy, 1999). Cela nous permet de faire directement le parallèle avec la constitution d'un marché mondial à travers la domination des plates-formes numériques. Google, Apple, Facebook et Amazon ont produit, avec l'aide des porte-conteneurs, le même effet d'échelle territoriale que la radio et le chemin de fer pour le territoire des marchés nationaux.

### 2.3. L'opinion publique existe, je l'ai mesurée

Le travail réalisé par Gallup pour le côté opérationnel (Gallup, 1939) et Lazarsfeld (Katz et Lazarsfeld, 1955) pour le côté scientifique n'est donc pas une simple opération marketing ou un *lifting* des sciences sociales : il fournit à des sociétés entières les méthodes pour s'auto-analyser, pour se représenter elles-mêmes comme opinions. Tarde avait beau avoir mis en évidence l'importance de ces opinions (Tarde, 1989), c'est seulement lorsque les métriques sont mises en place et produites de façon conventionnelle que l'opinion finit par exister. Et seules la commande des médias et leur capacité à produire de façon unifiée un public sur un territoire national permettaient de faire durer ce montage méthodologique. Le « tout » dont parlent les sondages, c'est en fait à l'origine le « public » constitué par les médias, qui permettent de faire émerger cette audience comme « opinion publique », de la rendre visible et mesurable en permanence. Cette parenté entre mesures d'audience et méthodes de suivi de l'opinion publique, parenté technique et historique, doit être considérée comme la clé du dispositif : les médias veulent avant tout mesurer des audiences ; c'est ce que fit Gallup pour la lecture, mais les techniques

mises en place se transformèrent en outils prédictifs de votes, ce qui justifia le pari sur une opinion publique. Le tout « audience » voire « public » a ainsi muté en « opinion publique » et a pu se détacher de son autoréférence aux médias – qui se mesuraient eux-mêmes au point d'être exploitables par les marques pour mesurer l'influence de leurs campagnes. Les « parties » que sont les expressions individuelles sont préformatées pour être enregistrables et calculables, mais le lien entre parties et tout (Latour *et al.*, 2012) n'est réalisé que par les boîtes noires des instituts de sondage. Les précautions scientifiques de rigueur sont prises grâce aux « intervalles de confiance » (définis en 1934 par Neyman), qui permettent de garder une référence avec l'exhaustivité de la population étudiée. À cet instant, chacun sait que « l'opinion existe », quel que soit le travail de compte rendu des artefacts nécessaires pour la faire exister et quoi qu'en dise Bourdieu[6]. Le travail de convention (Eymard-Duvernay *et al.*, 2004) ainsi réussi porte sur les mêmes assemblages de médiations déjà évoqués pour les traces :

- des *surveys* et des *polls* (à partir d'expressions individuelles cadrées par des questions et ainsi rendues calculables) ;
- assemblés et formatés par des instituts de sondage ;
- sous garantie de représentativité d'échantillons (*sampling*) ;
- pour des médias ;
- en vue d'un *monitoring* ;
- pour produire de l'opinion publique (et des audiences).

Comme le dit Alain Desrosières, l'essentiel n'est pas de savoir si ces données sont des reflets ou des miroirs de la société ou d'autre chose, mais de « faire quelque chose qui se tient » (Desrosières, 2001).

Notons qu'un élément nouveau intervient ainsi dans cette chaîne : celui de la contrainte méthodologique, exprimée en terme de représentativité des échantillons, car cet élément manque encore pour les traces numériques, ce qui explique en grande partie l'incertitude et la suspicion sur tous les résultats obtenus par comparaison avec les sondages, dont les « biais », sont bien connus mais contrôlés par convention depuis les années 1940. La « consolidation » qu'Emmanuel Didier décrit si bien pour les statistiques et les sondages hors études d'opinion, reste à faire.

Ce retour un peu long sur la fabrication réussie de l'opinion était nécessaire non seulement pour comprendre les analogies entre cette époque et celle où nous vivons, mais aussi pour mesurer le travail nécessaire pour produire des conventions de qualité équivalentes qui fassent exister « les traces » comme entités reconnues pour les sciences sociales. Il nous faut bien considérer l'opinion comme une réalité sociale qui vit sa vie et ne pose plus question grâce à la qualité des montages techniques et institutionnels qui ont stabilisé son mode d'apparition. Certes, le monde des sciences sociales, qui inclut la science politique, et celui du marketing restent bien séparés : pourtant, ils ont utilisé pendant des années les mêmes méthodes, voire les mêmes échantillons tout en étant capables de s'en distinguer. La question posée à ce nouveau monde des traces qui émerge sur le Web est du même type : comment pouvons-nous inventer les sciences sociales qui leur correspondent tout en admettant les conditions de production et d'utilisation de ces traces ?

6. « L'opinion publique n'existe pas », titre de 1984 qui introduisait un texte disant « l'opinion publique des sondages n'existe pas » (Bourdieu, 1984).

## 3. La fabrication de la « société »

Un autre moment historique des sciences sociales nous permettrait de complexifier le panorama et de le percevoir dans la longue durée. Nous prétendons en effet que Durkheim a réussi une opération identique à celle de Gallup et de Lazarsfeld, qui inventèrent l'« opinion publique », car il parvint à faire exister la « société ». Autant le caractère conventionnel de la notion d'opinion peut encore être admis, autant l'évidence de la société ne souffre pas discussion. D'autant que le terme ne date pas de Durkheim, même si son histoire n'est pas si longue. L'archéologie de la notion de société pourrait encore être enrichie par l'appel aux travaux de Quételet produisant son « homme moyen » (Quételet, 1846), qui resta longtemps la clé de toute la statistique. À la fin du XIX^e^ siècle cependant, et avec le coup de génie de Durkheim en grande partie, se produit un changement d'existence pour la notion de société. Les premiers travaux de Durkheim sur la division du travail social (Durkheim, 1893) ne s'appuyaient pas sur une méthode statistique mais posaient les bases d'un modèle de types sociaux agrégés en solidarités mécaniques et organiques. Avec *Le Suicide* (Durkheim, 1897), la méthode se met en place pour prolonger cette discussion des types qui va faire émerger l'anomie comme situation problématique. Mais l'appui sur les données produites par les États, consignées dans des registres issus de ses diverses composantes (ministères, préfectures, administrations), devient une clé dans la démonstration. Ce sont en effet ces agrégats qui sont expliqués ou explicatifs, grâce à une méthode de comparaison entre pays, entre régions, départements ou districts quand c'est possible et nécessaire. La méthode dépend entièrement des données disponibles et ne peut se payer le luxe de critiquer ou de mettre en doute les procédures de production de ces données, malgré les innombrables limites relevées dès la publication. En organisant tout son dispositif de preuve autour de ces statistiques administratives nationales, Durkheim trouve un analogue quantitatif à son parti-pris conceptuel qui place la « société » dans un statut à part de toutes les manifestations et de tous les comportements individuels. Le *tout* de Durkheim devient une entité de second degré (Latour, 2005), la « société », alors que les recensements et autres registres de données des États ne font pourtant qu'un travail de récupération d'événements administratifs individuels (état civil, procédures judiciaires, etc.), formatés dans des catégories identiques et agrégés pour faire apparaître des comportements de populations. Toute la force de conviction de Durkheim sera de faire exister ces populations statistiques comme équivalentes de sa « société ».

L'appareil statistique rend visible cette société de la même façon que le sondage rendra visible l'opinion et, dès lors, indépendamment de la validité statistique, le cadrage (*framing*) ainsi opéré gagne en puissance. Il faut en effet remarquer qu'une forme d'« alliance objective » se constitue entre les producteurs de données issus des administrations de l'État et les sciences sociales naissantes. Ensemble, ils vont produire l'entité « société » comme l'objet à suivre par l'État pour des raisons de gouvernement et à expliquer pour des raisons scientifiques. Le résultat tiendra dans une évidence partagée : la « société » existe, et les méthodes qui permettent de la faire exister n'ont pas lieu d'être interrogées puisqu'elles démontrent à la fois leur valeur scientifique et leur valeur opérationnelle, outil de preuve et outil de gouvernement comme le dit Desrosières (2014). Processus et alliances tout à fait identiques à celles que l'on rencontre entre les médias et les instituts de sondage qui s'entendent pour faire exister l'opinion et la rendre naturelle, la « considérer comme acquise » (*taken for granted*), après un long travail de montage de conventions.

### 3.1. Le temps des calculs et des machines à calculer

Dans le cas de Durkheim, il faut noter des voisinages historiques, qui ne valent pas causalité mais qui permettent de comprendre le gain de puissance de cette façon de faire exister la société. En effet, en 1890, Herman Hollerith utilise sa machine (qu'il a inventée quelques années auparavant et pour laquelle il a déposé une demande de brevet en 1886) pour réaliser le recensement américain. En effet, le Bureau of the Census n'avait pas réussi à finir de traiter le recensement précédent qui datait de 1880 lorsqu'il fallut déjà lancer le suivant. Un changement de technique était nécessaire et disponible. La machine de calcul mécanographique de Hollerith fit le travail et fut commercialisée pour les mêmes objectifs de recensement dans plusieurs pays, dont la France. La compagnie de Hollerith sera transformée par Watson en IBM en 1926. On comprend mieux comment la puissance gagnée dans le dénombrement et dans la description des populations consolide le statut de l'État et lui offre des sources de renseignements supposées utiles à son gouvernement. La prétention à l'exhaustivité du comptage accomplit la promesse du concept de société : les dispositifs techniques de saisie du tout existent, ce sont les machines de Hollerith équipant les procédures de recensement.

La performance de Durkheim aura ainsi été de faire tenir un assemblage de médiations fort puissant :

- des recensements ;
- assemblés et formatés par des administrations publiques ;
- sous garantie d'exhaustivité ;
- pour des États ;
- en vue d'un gouvernement ;
- pour produire de la « société » (à partir des populations) ;
- à l'aide de machines de calcul mécanographiques.

### 3.2. Le pouvoir d'agir des dispositifs techniques de calcul

Nous introduisons ici la dimension technique des supports de calcul qui produisent les données, car ces capacités de calcul et leur augmentation jouent un rôle essentiel. Les machines IBM qui servent les grands calculs des États vont irriguer toutes les institutions pendant 80 ans, et pénétrer de plus en plus profondément dans l'équipement de tous les services administratifs centraux puis locaux.

Peut-on trouver pareille situation pour l'invention de l'opinion publique ? Au moment même où Gallup adapte le *sampling* (échantillonnage) pour les sondages d'opinion et en fait la démonstration lors de l'élection de 1936, Alan Turing écrit son fameux article qui constituera les fondations de toute l'informatique (Turing, 1936). Avec John von Neumann, qui pensa quelques années plus tard l'architecture-type de l'ordinateur (Neumann, 1945), les conditions de développement de l'informatique et des calculs rapides émergent. Or, dans le cas de l'opinion publique, la perte de l'exhaustivité doit se compenser par un suivi plus fréquent et une réactivité plus importante nécessaire pour les médias. Seules les capacités des ordinateurs, associées à celles des réseaux téléphoniques pour la transmission des données, permettront à partir des années 1950 d'unifier et d'accélérer les calculs de ces échantillons représentatifs à une échelle nationale.

Dans la même veine, on mesure dès lors la mutation actuelle en cours avec Internet puis avec le Web. La fonction de suivi des traces telle que Google l'a pensée et équipée en 1998 dépend entièrement d'une architecture technique du Web inventée en 1990 par Berners-Lee et Cailliau. Dans ce cas, la dépendance technique est totale car il n'existe pas d'autres moyens de faire émerger ces liens entre sites, ces traces laissées par des clics et autres comportements des internautes. C'est aussi pour cela que l'assemblage entre les marques, les réseaux techniques et les traces est nettement plus fort que celui entre les médias, l'informatique et l'opinion, ou celui entre les États, le calcul mécanographique et la société.

## 4. Ce que les sciences sociales peuvent faire du numérique, ce que le numérique fait aux sciences sociales

Replacer les mutations numériques dans cette longue histoire des sciences sociales permet de mieux comprendre les mouvements contemporains dans l'usage des traces. Trois postures peuvent se présenter :

- l'une qui tente de reprendre le cours des sciences sociales des générations précédentes pour appliquer leurs méthodes et leurs concepts de « société » et d'« opinion » aux traces du Web ;
- une autre qui accepte ce nouveau monde des traces en s'immergeant dans ses exigences et ses principes en abandonnant les traditions et les impératifs scientifiques, et qui est résumée par l'argument de « la fin de la théorie » annoncée dans *Wired* par Chris Anderson (2008) et mis en effet en pratique dans les méthodes du Big Data.
- la dernière qui s'affronte à la radicale nouveauté de cette configuration socio-technique et qui tente de comprendre quelle peut être la place des sciences sociales dans la production de nouvelles conventions pour exploiter ces traces. Elle doit s'interdire de reprendre les concepts d'opinion et de société, et trouver un cadre conceptuel nouveau pour ces traces, qui valent pour elles-mêmes car elles ne sont plus générées que dans cet univers numérique. Nous présenterons ici les conditions de félicité d'une telle nouvelle génération sans nous étendre sur les formes d'exploitation des données web par les générations de la société ou de l'opinion, ni sur les conséquences d'une acceptation de la fin de la théorie par les sciences sociales.

### 4.1. Les propriétés des sciences sociales de troisième génération

La troisième génération de sciences sociales doit assumer le caractère radicalement nouveau de ces traces hétérogènes, sans les rabattre sur un statut de traces ou des symptômes d'un vrai social (la société, ou l'opinion), et sans pour autant se laisser happer par le système autoréférentiel de production/suivi des traces qui se dispense de théorie car il a d'autres visées. Nous présenterons d'abord ses propriétés générales, à la façon d'un cahier des charges, avant de revenir plus en détail sur les médiations jusqu'ici mobilisées et sur les choix qui s'imposent dans ce domaine.

Le mouvement orienté vers le Big Data peut fournir des premières pistes qui méritent d'être confrontées à celles des sciences sociales précédentes. Ainsi les critères de qualité du Big Data sont souvent résumés aux 3V : volume, variété, vélocité. La parenté avec les exigences des sciences sociales est assez frappante.

### 4.1.1. Volume et exhaustivité

Le volume correspond à l'exigence d'exhaustivité traduite sous un mode quelque peu limité, puisque sur le Web, rien ni personne ne permet de définir les frontières des univers de données rassemblées. Dès lors, il conviendra de fixer un équivalent de ce volume qui se rapproche des exigences traditionnelles de l'exhaustivité, sans pour autant pouvoir les suivre lorsqu'on traite du Web. Les protocoles de constitution de corpus de données pourraient ainsi être normalisés pour assurer qu'un volume suffisant soit atteint et justifiable. Le problème est en général simplifié dans l'approche du Big Data, dans la mesure où les volumes sont aisément accessibles mais rarement justifiables. Comme toute méthode, l'enjeu n'est pas de fixer des standards *a priori*, qui seraient rapidement dépassés en raison de l'évolution rapide des volumes produits[7]. Il s'agit plutôt de fixer les éléments de référence qui permettent de définir des seuils suffisants, l'important étant dans cette prudence aristotélicienne à définir *ce qui convient*, pour un travail scientifique, et non plus pour les traitements opérationnels déjà évoqués dans les postures précédentes. Nous devons clairement faire notre deuil de l'exhaustivité, mais cela ne dispense pas de fixer les cadres conventionnels de toute démarche en sciences sociales traitant de traces numériques.

### 4.1.2. Variété et représentativité

Le deuxième critère, la variété, est lui aussi une forme de transcription de l'exigence de représentativité qui a permis à toutes les sciences sociales de procéder par enquêtes, par sondages, à base d'échantillonnage. Là encore, le critère est une version lâche de la représentativité, qui suppose que l'on accepte un niveau *suffisant* de variété. Tout chercheur qualitatif en sciences sociales avait à cœur de s'assurer de cette variété suffisante, pour des buts de comparaison ou pour assurer seulement qu'il ne ciblait pas un groupe trop particulier. Parfois, cependant, cette méthode devait être contestée lorsqu'il était nécessaire d'aller observer des groupes atypiques ou dysfonctionnels par exemple pour faire apparaître des phénomènes qui, sous l'effet des lois normales, auraient disparu. La variété devient alors un critère qui permet d'aller chercher non pas des moyennes, mais des extrêmes, ce qui se fait en clinique sociologique. La variété dans le Big Data peut prendre des dimensions très différentes selon les contextes. 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. Nous introduisons ici un autre élément qui doit rester une clé dans le travail de convention à produire pour les sciences sociales de troisième génération : aucune description du « social-société », du « social-opinion » ou du « social-traces » ne peut plus être produite en généralité. La prolifération

7. De 1,2 zettaoctets en 2010 à 2,8 zettaoctets en 2012 [source : International Data Corporation] pour ce qui est considéré par le Big Data, 1 Mo = 10 puissance 6, un Zo = 10 puissance 21.

des traces rend paradoxalement 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*, 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 forme de représentativité et d'exhaustivité. En effet, sous ces conditions de limitation à des *issues* (Marres, 2007 ; Marres et Weltevrede, 2013), il devient possible de rendre compte non seulement de propagations, de flux, mais aussi de stabilisations et d'alliances, qui constituent du « social-toujours-local », quand bien même il s'agit de traiter d'*issues* que l'on qualifie habituellement de macro, de grandes échelles ou de longues durées. Comme on le voit, l'approche de l'acteur-réseau et de la traduction (Akrich, Callon et Latour, 2006) constitue une ressource particulièrement adaptée pour traiter de phénomènes numériques sans y projeter des catégories existantes.

### 4.1.3. Vélocité et traçabilité

Enfin, le dernier critère, la vélocité, ne trouve guère d'équivalent dans les sciences sociales de première et de deuxième génération. Cela a ouvert un espace pour l'étude de certains phénomènes sociaux par des disciplines hors des sciences sociales qui ont produit des modèles qui sont toujours appliqués à certains phénomènes comme ceux de la ségrégation sociale (modèles de Schelling, développant le point de basculement ou *tipping point*) ou encore à l'étude de la *ola* (Farkas *et al.*, 2002) et d'autres phénomènes de foule dans les lieux publics (Theraulaz et Benabeau, 1999). Comme on le voit, ce sont certains types de processus sociaux qui peuvent être pris en compte, non le suivi d'une discussion sur le Web et sa transformation dans le cours de l'action collective, dans le cadre d'une controverse par exemple. Il ne nous semble pas possible de fonder des sciences sociales de troisième génération sur des modélisations, certes puissantes, d'objets aussi simplifiés, mais il faut reconnaître que ces travaux d'une part signalent un type de phénomènes qui relèvent sans doute de ce niveau d'analyse (les mouvements de foule par exemple) (Boullier, 2010), et d'autre part mobilisent des modèles qui peuvent donner des pistes pour l'analyse de toute propagation.

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 *memes* qui se propagent sur le Web (comme les images animées en GIF en sont devenues les prototypes) (Shifman, 2014). Il est très significatif que Jon Kleinberg, celui-là même qui avait exporté les méthodes de la scientométrie (Courtial *et al.*, 1993) vers l'étude de la topologie du Web et qui fut repris par Google, s'est intéressé depuis plusieurs années (Kleinberg, 2002) à la mise au point d'un *meme tracker* avec Leskovec (Leskovec *et al.*, 2009). Leur étude la plus fameuse a porté sur la propagation des citations durant la campagne de Barack Obama, ce qui leur permit de réaliser une visualisation spectaculaire de la focalisation de l'attention en courbes à montée et à descente très rapides (*streams and cascades*) autour de certains incidents de la campagne. Leur méthode agrège tous les types de traces que peuvent laisser ces citations, traitées comme des chaînes de caractères dont on peut trouver la trace dans tout le Web, et en produit une métrique ancrée dans le temps, au jour le jour, voire minute par minute désormais avec Twitter (l'unité de mesure étant devenue le

*tweet per second* ou TPS). Cette approche par les *memes* peut nous inspirer sous deux réserves (au-delà des réserves pour l'idéologie de Richard Dawkins (1976), fondateur de la mémétique) :

- Il faudra la rendre capable de suivre les transformations-traductions de ces *memes* dans des milieux différents, car « toute existence va différant » comme le disait Tarde, et que l'imitation qu'il a si bien mise en avant était selon lui un processus permanent couplé à l'opposition qui générait tout autant des hésitations et dès lors des adaptations.
- Il faut admettre non seulement de suivre des *issues* comme nous l'avons dit précédemment mais des entités circulantes, qui chez Tarde (2001) se classaient en deux ensembles, les croyances et les désirs. Ce qui veut dire suivre la trace des transformations de ces *memes* les plus élémentaires, non pas pour leur donner un statut d'atome mais pour repérer leur pouvoir d'agrégation, de propagation et de transformation, comme autant de médiations qui tissent le social de la troisième génération. En ce sens, il convient d'abandonner toute référence à des acteurs au sens de « société » ou d'« opinion » (c'est-à-dire des sujets humains) et de prendre en compte la puissance d'agir de toute entité, son *agency*.

Dès lors, et à ces conditions, il devient possible de trouver un équivalent de la vélocité du Big Data. Certes, la vélocité intéresse tout processus de propagation et permet de trouver des *patterns* de flux par exemple. Cependant, l'objet n'est pas ici une mécanique des flux mais bien le statut des entités sociales qui sont nécessaires à fonder des sciences sociales de troisième génération. Nous dirons donc qu'il convient de considérer la *traçabilité* comme le critère essentiel de qualité des entités que l'on peut étudier et qu'il sera nécessaire de produire les conditions de félicité de l'étude de cette traçabilité. Nous pouvons en donner quelques-unes à titre d'exemples.

- 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.
- Les traces en question doivent permettre des suivis d'associations hétérogènes, ce que nous pourrions appeler 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 et à suivi.
- Le suivi des traces en question doit permettre de *dater* avec précision tous les événements, toutes les transformations et toutes les associations. S'il est en effet possible d'accepter de perdre le critère de représentativité et d'exhaustivité, il devient totalement inutile de repérer des agrégats de traces dont on aurait perdu précisément la traçabilité, et dont on ne saurait rien de leur état précédent.

De toutes ces conditions à l'établissement d'une convention de traçabilité, il faut retenir que l'on s'intéresse aux pouvoirs d'association parmi lesquels la propagation est la plus significative de façon à dégager des trajectoires plutôt que des positions.

### 4.2. Conventions académiques et conventions des plates-formes

Comment parvenir à produire la convention qui ferait tenir une science des traces ? Les acteurs essentiels de ces traces sont les plates-formes – GAFA (Google, Apple, Facebook, Amazon) en résumé –, mais aussi les marques qui font vivre tout cet écosystème par l'exploitation publicitaire de cette traçabilité. De même que la sociologie durkheimienne s'est associée de fait avec les institutions étatiques productrices de données pour produire la « société » en combinant de fait enquêtes (des statisticiens) et registres (des administrations) (Desrosières,

2008), les instituts de sondage de Gallup et de Lazarsfeld se sont associés aux médias, grands consommateurs de données sur les publics, pour produire leur « opinion publique », en rapprochant ainsi « mesures d'audience » (le public des médias) et « opinion publique » (le public de la sphère publique au sens politique). Les sciences sociales de troisième génération ne pourront guère faire autrement que de s'associer à ces plates-formes et à ces marques pour produire la science des traces qu'il est possible aujourd'hui d'imaginer, dans des termes somme toute pas si éloignés de ceux que Tarde avait annoncés. Mais, comme nous l'avons vu, il est aussi possible de s'associer à ces parties prenantes du monde des traces pour renforcer les sociologies de la société et de l'opinion, ce qui reste tout-à-fait légitime et nécessaire, mais qui peut devenir très dangereux si les principes n'en sont pas posés indépendamment des opérateurs en question. L'enjeu que nous soulevons ici consiste donc à inventer les conventions qui feront tenir la sociologie de la troisième génération, celle qui prend en compte la perte de l'exhaustivité et de la représentativité, pour les réinventer en volume, en variété et en traçabilité. Plusieurs procédures d'invention de ces conventions sont possibles :

- Les traces produites étant dépendantes des plates-formes, on pourrait espérer les *modifier à la source* et obtenir ainsi une forme d'accord similaire à celui qui s'est mis en place entre les instituts de sondage et les chercheurs académiques, ces derniers réutilisant les échantillons, les techniques, voire les enquêtes pour un traitement secondaire plus approfondi.
- La seconde stratégie consiste à exploiter les traces produites par les plates-formes en les *détournant* de l'usage pour lequel elles avaient été conçues (*repurposing*, Rogers, 2013). La propagation est en elle-même suffisamment intéressante pour générer un volet permanent des sciences des traces.
- La troisième stratégie consiste à produire le cadre conceptuel qui permettra de constituer les objets scientifiques issus de l'exploitation des traces, objets propres aux sciences sociales et non réductibles à l'usage fait par les marques. Aux couples registre / enquête, 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 *réplications*. Le terme présente une parenté avec les *memes* ou « mèmes » (la mémétique considère le *meme* comme un réplicateur) car les traces nous intéressent pour suivre des réplications (*replicas*), des imitations au sens tardien (et donc des oppositions et des adaptations). Il est aussi apparenté au terme *répliques* issu du monde de l'échange langagier pour désigner des réparties dans un dialogue (*replies*), dans une conversation (Boullier, 2004a et 2004b), qui aurait dû être au centre des sciences sociales selon Tarde. Il permet ensuite de filer une métaphore suggestive avec les réplications des tremblements de terre (*aftershock*).

L'intérêt principal de ce terme tient dans le décentrement réalisé vis-à-vis des notions de structures (société), d'acteurs (opinion), 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 réplications. 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 avec les plates-formes mais selon nos objectifs.

Le principe d'une sociologie des réplications repose sur l'impératif de suivre des éléments, sans pour autant savoir comment ils vont s'agréger pour faire des « tout » à géométrie variable. 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 la théorie de l'acteur-réseau (Akrich *et al.*, *op. cit.*). L'objet d'étude n'est pas tant l'élément, qui peut avoir des attributs très variés, en étendue et en matière, ni seulement les agrégats, ce que l'on tend à faire avec les clusterisations, mais bien le processus de circulation et d'agrégation ou de désagrégation, au moment de bifurcation des courbes. Dans ces courbes, il faut alors plutôt se focaliser sur les moments d'émergence et d'évanescence, et non sur les pics qui fonctionnent comme des agrégats, comme le fait le *memetracker* de première génération. L'objet de cette science des réplications est bien l'agentivité des réplications qui se propagent et qui finissent par nous prendre. Car les individus sont en fait traversés par les idées, et les idées « nous agissent », non l'inverse, comme l'avait bien indiqué Tarde. « 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 » (Latour, 2011). Il est alors possible d'étudier les propriétés de ces réplications pour comparer éventuellement leurs chances de survie ou de contamination rendues possibles par ces différences de propriétés. Comme on le voit, l'approche par les réplications est alors une entrée vers une monadologie (Tarde, 1893) – qui se différencie radicalement d'une vision atomiste. Nous avons commencé à le faire sans l'appareillage statistique dans le cas des tags de photos de la base de données Flickr (Boullier et Crépel, 2013) en montrant le pouvoir de connexion d'un tag « bras croisés » sur une photo de Savorgnan de Brazza commentée par Roland Barthes : le *punctum* circulait mieux que le *studium*, et produisait de nouvelles relations. Mais les travaux réalisés sur les n-grams étudiés à partir de Google Books (Michel *et al.*, 2011) ont permis de montrer des évolutions de la langue anglaise (le prétérit des verbes irréguliers). Lev Manovich (2012) a constitué une base de mangas de plus d'un million d'exemplaires pour comparer les attributs les plus élémentaires, comme le contraste, et produire une vision inédite des influences entre courants, et il a exploité des outils de similarité identiques pour réaliser des comparaisons culturelles entre pays à partir de millions de photos sur Instagram ou sur la place Maidan à Kiev. Mariannig Le Béchec (Le Béchec et Boullier, 2014) a recensé les vignettes des drapeaux bretons sur les sites qui affichent un lien avec la région pour montrer comment une telle réplication, qu'elle nomme « signe transposable », circule bien au-delà d'un territoire. Tous ces exemples exploitent certaines des propriétés de ces traces – variété, volume ou vélocité – dans des proportions différentes. Notons qu'aucun ne se soucie d'expliquer les caractéristiques de ces propagations par des causes (externes ou internes, qui seraient « plus sociales ») et qu'ils en font seulement l'inventaire, qu'ils les suivent, pour rendre compte de leur pouvoir de circulation propre.

### 4.3. L'extension du domaine des traces

L'ère des traces ne fait que commencer cependant, 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. 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 telle qu'elle est annoncée. Il ne sera plus possible de renvoyer à des personnes, à des entités sociales au sens des sciences sociales de première et de deuxième génération. De plus, il n'y a pas de raison pour que les sciences sociales ne s'emparent pas de ces nouvelles sources. La théorie de l'acteur-réseau et toutes les approches qui ont pris en compte la matérialité des échanges (ex : cognition distribuée, située, théorie du support, médiologie, etc.) et l'interobjectivité (Latour, 1994) ne seront pas surprises par cette nécessaire prise en compte d'entités matérielles équipées de capteurs, d'effecteurs, de traceurs, etc. C'est pour cette raison aussi que les sciences sociales qui traitent de ces traces ne peuvent plus s'appuyer sur cette définition restreinte du social qui a présidé à leur création. Dans cet Internet des objets en effet, aucune garantie sur le statut sociologique des entités ne peut être recherchée, aucun état civil ne permet de préorganiser un garant indiscutable comme c'est le cas pour les personnes.

## Conclusion

Il est nécessaire de fournir un tableau synthétique des trois âges des sciences sociales qui aura l'avantage de rendre perceptible la cohérence de l'approche, mais qui oblige dans le même temps à schématiser et à éliminer des spécificités propres à chaque âge.

| | **1re génération** | **2e génération** | **3e génération** |
|---|---|---|---|
| **Concept du social** | Société(s) | Opinion(s) | Réplication(s) |
| **Dispositifs de collecte** | Recensement | Sondage | Traces (*big data*) |
| **Principe de validation** | Exhaustivité | Représentativité | Traçabilité |
| **Co-construction institutions/recherche** | Registre/enquête | Audience/sondage | Traces / *big data* |
| **Acteurs majeurs de référence (et financeurs)** | États | Médias de masse / *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** | Division du travail et État-providence | Propagande et influence des médias (mesures d'audience) | Science et technologie (scientométrie) |
| **Conjoncture technique** | Machines de Hollerith (calcul mécano-graphique) | Téléphone, radio puis informatique | Internet, Web et *big data* |
| **Formats sémiotiques** | Tableaux croisés et cartes topographiques | Courbes et histogrammes / diagrammes circu-laires (camemberts) | Graphes, *timelines, dashboards* |

| **Métriques** | Statistique fréquentiste | *Sampling* | Similarités (patterns) |
|---|---|---|---|
| **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 |

**Tableau 1.** Les trois générations de sciences sociales.

La cohérence toujours abusive du tableau ne doit pas faire oublier que ce qui est en jeu est la construction d'une offre de sciences sociales de troisième génération qui n'est pas garantie. La tendance à la fin de la théorie et l'occupation du terrain par les plates-formes du Web (GAFA) qui produisent, calculent et publient sur ces traces elles-mêmes reste dominante, et cela pour des visées commerciales avant tout, puisque les marques sont les grands demandeurs de ces approches. Cela n'invalide ni l'intérêt pour les marques d'apprendre à réagir en utilisant ces métriques, ni le rôle des sciences sociales de la société et de l'opinion de continuer à développer leurs approches en utilisant ces sources. Notre intention est seulement de contribuer à poser les bases d'une convention permettant de faire émerger une théorie sociale et un objet, les réplications, qui ne rabattent pas le numérique sur les « méthodes numériques » ni sur les « humanités numériques ». Il existe une nouvelle matière première qui mérite un examen pour elle-même et qui produit une troisième couche 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 réplications doivent finir par exister au même titre.

## Références


Anderson C. (2008), « The end of theory: the data deluge makes the scientific method obsolete », *Wired*, 24 juin 2008.

Akrich M., Callon M. et Latour B. (2006), *Sociologie de la traduction. Textes fondateurs*, Paris, Presses des mines de Paris.

Blondiaux L. (1998), *La Fabrique de l'opinion. Une histoire sociale des sondages*, Paris, Seuil. Plusieurs remarques prennent appui sur ce travail fondamental pour notre réflexion.

Boullier D. et Lohard A. (2012), *Opinion mining et sentiment analysis. Méthodes et outils*, Paris, OpenEdition Press (http://books.openedition.org/oep/198).

Boullier D. (2013), « 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, Karthala, 492 p.

Boullier D. et Lohard A. (2013), « Médiologie des réputations », *Journées d'étude Association française de sociologie : vers une sociologie des réputations ?*, Amiens.

Boullier D. (2004a), *La Télévision telle qu'on la parle. Trois études ethnométhodologiques*, Paris, L'Harmattan.

Boullier D. (2004b), « La fabrique de l'opinion publique dans les conversations télé », *Réseaux*, n° 126, p. 57-87.

Boullier D. et Crepel M. (2013), « Biographie d'une photo numérique et pouvoir des tags : classer/circuler », *Revue d'anthropologie des connaissances*, vol. 7, n° 4, p. 785-813.

Boullier D. (2015a), « Vie et mort des sciences sociales avec le Big Data », *Socio*, n° 4, p. 19-37 (https://socio.revues.org/1259).

Boullier D. (2015b), « Les sciences sociales face aux traces du Big Data : société, opinion ou vibrations ? », *Revue française de science politique*, vol. 65, n° 5, p. 805-828.

Bourdieu P. (1984), « L'opinion publique n'existe pas », *Questions de sociologie*, Paris, Minuit, p. 222-235.

Boyd D. et Crawford K. (2011), « Six provocations for *big data* », *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, Oxford Internet Institute / University of Oxford, Oxford, UK.

Cardon D. (2013), « Du lien au *like* sur Internet. Deux mesures de la réputation », *Communications*, n° 93 (*La Réputation*), p. 173-186 (https://www.cairn.info/revue-communications-2013-2-page-173.htm).

Cochoy F. (1999), *Une histoire du marketing. Discipliner l'économie de marché*, Paris, La Découverte.

Courtial J.-P., Callon M. et Penan H. (1993), *La Scientométrie*, Paris, Presses universitaires de France.

Dawkins R. (1976), *The Selfish Gene*, Oxford, Oxford University Press.

Desrosières A. (1993), *La Politique des grands nombres. Histoire de la raison statistique*, Paris, La Découverte.

Desrosières A. (2001), « Histoire de la raison statistique : le moment bayésien », *Courrier des statistiques*, n° 100.

Desrosières A. (2008), *Gouverner par les nombres. L'Argument statistique II*, Paris, Presses de l'École des mines (http://books.openedition.org/pressesmines/341).

Desrosières A. (2014), *Prouver et gouverner : une analyse politique des statistiques publiques*, La Découverte, 284 p. Recueil posthume de textes choisis et rassemblés par Emmanuel Didier.

Didier E. (2009), *En quoi consiste l'Amérique ? Les statistiques, le New Deal et la démocratie*, Paris, La Découverte.

Driscoll K. et S. Walker (2014), « Working within a black box: transparency in the collection and production of big Twitter data », *International Journal of Communication*, nº 8, p. 1745-1764.

Durkheim E. (1897), *Le Suicide*, Paris, Alcan.

Durkheim E. (1893), *De la division sociale du travail*, thèse présentée à la Faculté des lettres de Paris, Paris, Alcan.

Eymard-Duvernay F., Favereau O., Orléan A., Salais R. et Thevenot L. (2004), « L'économie des conventions ou le temps de la réunification dans les sciences sociales », *Problèmes économiques*, nº 2838, La Documentation française, Paris.

Farkas I., Helbing D. et Vicsek T. (2002), « Social behaviour: Mexican waves in a excitable medium », *Nature*, vol. 419, nº 6903, p. 131-132 (https://www.nature.com/nature/journal/v419/n6903/full/419131a.html).

Gallup G. (1939), *Public Opinion in a Democracy*, Herbert L. Baker Foundation, Stafford Little lectures.

Katz E. et Lazarsfeld P. (1955), *Personal Influence: The Part Played by the People in the Flow of Mass Communication*, Glencoe, Free Press.

Kleinberg J., Gibson D. et Raghavan P. (1998), « Inferring web communities from link topology », in *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (HYPER-98)*, New York, p. 225-234.

Kleinberg J. (2002), « Bursty and hierarchical structure in streams », *Proceedings of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*.

Latour B., Jensen B., Venturini T., Grauwin S. et Boullier D. (2012), « The whole is always smaller than its parts'. A digital test of Gabriel Tarde's monads », *British Journal of Sociology*, vol. 63, nº 4, p. 590-615.

Latour B. (2005), *Reassembling the Social – An Introduction to Actor-Network-Theory*, Oxford, Oxford University Press; traduction française : Latour B. (2006), *Changer la société. Refaire de la sociologie*, Paris, La Découverte.

Latour B. (2011), « Gabriel Tarde. La société comme possession. La preuve par l'orchestre », *in* Debaise D., *Philosophie des possessions*, Les Presses du réel.

Latour B. (1994), « Une sociologie sans objet ? Remarques sur l'interobjectivité », *Sociologie du travail*, n° 4, p. 587-607.

Le Béchec M. et Boullier D. (2014), « Communautés imaginées et signes transposables sur un "web territorial" », *Études de communication*, nº 42, p. 113-125 (http://www.cairn.info/revue-etudes-de-communication-2014-1-page-113.htm).

Leskovec J., Backstrom L. et Kleinberg J. (2009), « Meme-tracking and the dynamics of the news cycle », *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD).

Manovich L. (2012), « Media visualization: visual techniques for exploring large media collections », *in* Gates K. (éd.), *Media Studies Futures*, Blackwell.

Marres N. (2007), « The issues deserve more credit: pragmatist contributions to the study of public involvement in controversy », *Social Studies of Science*, n° 37, p. 759-778.

Marres N. et Weltevrede E. (2013), « Scraping the social? Issues in live social research », *Journal of Cultural Economy*, vol. 6, n° 3, p. 313-335.

Michel J.-B., Shen Y.K., Aiden A.P., Veres A., Gray M.K. *et al.* (2011), « Quantitative analysis of culture using millions of digitized books », *Science*, vol. 331, n° 6014 [publié en ligne avant l'édition imprimée : 12/16/2010] (http://science.sciencemag.org/content/331/6014/176).

Neumann J. von (1945), *First Draft of a Report on the EDVAC*.

Orléan A. (2011), *L'Empire de la monnaie. Refonder l'économie*, Seuil, Paris.

Quételet A. (1846), *Lettre à S.A.R. le Duc régnant de Saxe Cabourg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politiques*, Bruxelles, Hayez.

Rogers R. (2013), *Digital Methods*, Cambridge (Mass.), MIT Press.

Shifman L. (2014), *Memes in Digital Culture*, Cambridge, MIT Press.

Tarde G. (1893), *Monadologie et sociologie*, Paris, Alcan, 55 p.

Tarde G. (1989), *L'Opinion et la Foule* [1901], Paris, PUF, 1989.

Tarde G. (2001), *Les Lois de l'imitation* [1895], Paris, Les Empêcheurs de penser en rond.

Turing A. (1936), « On computable numbers, with an application to the Entscheidungsproblem », *Proceedings of the London Mathematical Society*, vol. 42, n° 2, p. 230-265.

Zask J. (2000), *L'Opinion publique et son double. Livre I : L'opinion sondée. Livre II : John Dewey, philosophe du public*, Paris, L'Harmattan.