Conférences du Collège de France

Big data et traçabilité numérique

Les sciences sociales face à la quantification massive des individus

sous la direction de
Pierre-Michel Menger
et Simon Paye

COLLÈGE DE FRANCE
1530

Big data et traçabilité numérique

Les sciences sociales face à la quantification massive des individus

Conférences du Collège de France

Big data et traçabilité numérique

Les sciences sociales face à la quantification massive des individus

sous la direction de

Pierre-Michel Menger et Simon Paye

avec les contributions de

Jean-Samuel Beuscart, Dominique Boullier, Franck Cochoy, Éric Dagiral, Jérôme Denis, Samuel Goëta, Bernard E. Harcourt, Pierre-Michel Menger, Sylvain Parasie, Simon Paye, David Pontille, Guillaume Tiffon, Didier Torny, Jean-Sébastien Vayre

COLLÈGE
DE FRANCE
1530

Conférences du Collège de France

La vie scientifique et intellectuelle du Collège de France s'étend au-delà de l'enseignement qui y est prodigué. De nombreux colloques internationaux, séminaires de recherche et conférences de professeurs étrangers sont organisés chaque année. Et au sein des chaires et des laboratoires, plusieurs centaines de chercheurs engagent des travaux novateurs. La collection « Conférences du Collège de France » a vocation à refléter cette activité. Nativement numérique, publiée en accès ouvert freemium sur OpenEdition Books (https://books.openedition.org/cdf/1419), elle paraît également désormais sous forme imprimée.

Cet ouvrage a été réalisé avec la chaîne d'édition structurée XML-TEI Métopes développée par le pôle Document numérique de la MRSH de Caen.

Maquette : Mona Vallery

Éditeur : Collège de France

Collège de France / Publications
11, place Marcelin-Berthelot
75231 Paris Cedex 05

L'édition électronique de cet ouvrage est disponible à l'adresse suivante :
https://books.openedition.org/cdf/4987

Sommaire

Introduction

Pierre-Michel Menger
Professeur au Collège de France, titulaire de la chaire « Sociologie du travail créateur »

La production en masse de données numériques a rapidement ouvert la possibilité d'exploiter des informations en quantités inédites[1]. Un nombre croissant d'activités humaines, des plus privées aux plus publiques et relationnelles, peuvent désormais être analysées avec des moyens nouveaux dont la puissance d'exploration augmente très rapidement. Les traces numériques laissées par les usagers des moteurs de recherche, des réseaux sociaux ou des sites d'achat en ligne engendrent une masse d'informations sans précédent, non seulement sur les usages d'Internet en général, mais aussi sur les centres d'intérêt des individus, leurs pratiques de consommation et leurs orientations politiques ou religieuses. Au-delà d'Internet, une production comparable de données numériques de masse est assurée par un nombre croissant d'objets connectés qui peuvent être vus comme autant de capteurs des activités humaines (systèmes de géolocalisation, relevés de consommation d'électricité, systèmes de mobilité urbaine, cartes bancaires, passeports biométriques, etc.).

Depuis quelques années, l'expression « *big data* » (qu'on peut traduire par « données numériques massives ») sert à désigner ce phénomène. Elle tend à envelopper dans un même ensemble toutes les données numériques issues de l'activité des individus, qu'il s'agisse d'actes privés de consommation ou d'interventions dans les réseaux sociaux, ou de la production d'informations, de documents, d'archives dans les entreprises et les administrations, ou encore de la numérisation de tout le stock des données archivées par ces mêmes firmes, administrations, bibliothèques, et sources d'information.

Une nouvelle économie se structure autour d'un flux rapide d'innovations et d'applications destinées à stocker, à analyser, à visualiser et à échanger ces données, *via* des transactions marchandes ou hors marché. La recherche scientifique fondamentale et appliquée, l'innovation industrielle, le commerce, le marketing, l'expérimentation sociale, les politiques publiques sont invitées à se soucier de l'extraction et de l'exploitation de ce nouveau gisement informationnel. Les échelles d'analyse sont spectaculairement emboîtables depuis les cartographies planétaires d'échange et de diffusion d'informations jusqu'à la microsegmentation des comportements et des préférences des individus. Les unités d'analyse – la personne, l'organisation, le réseau, le groupe, la masse – sont aussi rendues beaucoup plus directement commensurables.

Les outils d'exploration de ces données bousculent non seulement l'analyse des comportements, mais encore les techniques d'influence et d'incitation. La détection à grande échelle de corrélations, la recherche automatisée de patterns, les algorithmes d'analyse sémantique prennent le pas sur les patients travaux d'enquête par échantillonnage et rivalisent avec

1. Je remercie chaleureusement Simon Paye pour ses remarques sur cette introduction et pour l'ensemble du travail accompli pour organiser le colloque qui s'est tenu au Collège de France et la production de cet ouvrage, qui en est issu.

l'expérimentation contrôlée. Profiler pour prédire les tendances et les usages, pour persuader les usagers et les consommateurs, et, ce faisant, pour obtenir que les anticipations issues des profilages se vérifient, voilà qui introduit de multiples boucles dans l'interaction avec les systèmes, boucles qui conduisent à ce dont le marketing a toujours rêvé : inventer une technologie silencieuse de la prophétie ou de la persuasion auto-réalisatrice, qui se substitue au bruit de la persuasion publicitaire traditionnelle.

Les entreprises numériques se multiplient et viennent nourrir les hypothèses de la recherche en sciences sociales. Lorsque les machines et les algorithmes apprennent progressivement à se passer d'instructions pour trouver, décider, exécuter, la création des métiers appariés à ces innovations obéit à la dynamique de concurrence des professions : de nouveaux savoirs et de nouvelles compétences, de nouveaux outils et de nouveaux services déplacent et redistribuent les aires d'expertise tout autant que la hiérarchie des problèmes, des méthodes et des chances de gain. Dans la production du savoir, une nouvelle génération de relations émerge entre la recherche fondamentale et la recherche appliquée, comme entre le monde académique et le monde industriel. Stanford et la Californie attirent un nombre croissant d'universitaires et de chercheurs. Cette agglomération de talents est multiculturelle, et ses comportements sont désinhibés : l'*academic entrepreneurialism* faisait parfois l'objet d'âpres controverses à la fin du XX^e^ siècle (Slaughter et Leslie, 1997 ; McSherry, 2001 ; Bok, 2003 ; Popp Berman, 2012), mais les réticences à son égard sont aujourd'hui levées face à l'effervescence de l'innovation, à la densité des échanges disciplinaires et à l'affaiblissement des hiérarchies de légitimité qui plaçaient la science fondamentale très au-dessus de la science appliquée, et les disciplines nobles, productrices de théories et de modèles, très au-dessus des disciplines ancillaires, immergées dans l'exploration des données et dans l'amélioration incessante des technologies. Ces chercheurs veulent tirer parti de l'agglomération des entreprises de haute technologie, de la capitalisation boursière considérable des champions de l'économie numérique, mais aussi d'un éthos de la démesure, de l'audace sans limites, qui s'établit tranquillement comme une nouvelle norme. La finitude de la vie ? Parions de pouvoir dessiner sans trop tarder l'asymptote de la vie éternelle. Le défi climatique ? Inventons la ville verte, qui s'autocontrôle, et la fluidité du trafic routier, que promet la voiture intelligente et autonome. Le déplacement dans l'espace ? Un peuple d'automates deviennent agents livreurs (les drones) ou chauffeurs (encore les voitures intelligentes), et le monopole des agences spatiales publiques s'achève.

De nombreux travaux sont consacrés depuis près de vingt ans à la polarisation du marché du travail et à l'affaissement du continent des emplois de qualification moyenne. Les relations de complémentarité ou de substitution entre les emplois et les technologies numériques expliquent pourquoi les tâches à forte intensité cognitive et créative (Autor, 2013 ; Autor et Handel, 2013 ; Brynjolfsson et McAfee, 2014) et à fort pouvoir relationnel (Deming, 2015) protègent les emplois supérieurs peu routiniers de la menace de robotisation à brève échéance, et pourquoi les tâches routinières, cognitives comme manuelles, sont exposées directement à la substitution du capital (robot, machines intelligentes, algorithmes) au travail. L'éthos de la créativité et de l'inventivité transgressives, qui suscite tant de fascination et de sidération, incarne en quelque sorte la nouvelle norme comportementale des humains non robotisables.

De même que le « principe des 4 V » (volume, variété, vélocité, véracité) est la signature de la séduction exercée par les données numériques de masse, selon Serge Abiteboul (2012), la croissance est spectaculaire dans le volume des échanges, des expérimentations, des innovations, mais aussi des débats, interrogations et critiques qui touchent aux dimensions tout à la fois technologiques, industrielles, entrepreneuriales, sociales et éthiques des transformations en cours. Les marchés, la recherche-développement, les sociétés savantes, les législateurs, les organisations publiques internationales, les communautés de recherche, les associations de consommateurs, les médias ont, chacun depuis sa perspective, identifié un ensemble de questionnements communs, dont le pouvoir de rayonnement et de contagion illustre à lui seul la puissance des mécanismes de production et d'échange des contenus informationnels (Mayer-Schönberger et Cukier, 2013). Les débats ainsi lancés sont d'autant plus intenses que le maniement des données massives 1) s'applique à tous les aspects de la vie sociale et économique, 2) bouscule tous les partages habituels entre privé et public, entre intime et relationnel, entre secret et révélation, entre consentement passif et consentement explicite, entre usage et confiance, et 3) fonde son expansion sur la nature intrinsèquement relationnelle des données, qui décloisonne et fait communiquer toutes les dimensions et tous les domaines d'activité générateurs d'informations.

De toutes les propriétés mises en avant par les acteurs promoteurs des « *big data* », le volume et la croissance exponentielle de ces masses paraissent être les plus stupéfiantes. Pourtant, nous sommes devenus familiers des grandeurs extrêmes : pour connaître les bornes de notre monde, nous recourons, par exemple, à des ordres de grandeur dont l'échelle varie de 5×10^{-324} (en informatique, c'est la plus petite valeur positive différente de zéro qui peut être représentée par une valeur à virgule flottante à double précision) à 10^{120} (quantité de toutes les différentes parties d'échecs qu'il est possible de jouer, dit « nombre de Shannon »). S'agissant des « *big data* », c'est en réalité la capacité de mise en relation ou de connectivité de ces données qui est le phénomène le plus remarquable. N'importe quelle parcelle de réalité peut être considérée comme une somme d'informations, informations dont la valeur vient de leur mise en relation. Et n'importe quelle relation observable et traçable entre un individu et un élément de son environnement peut être génératrice de données enregistrables, stockables, communicables et échangeables.

Le mouvement identifié sous le vocable « Big Data » est double. Il fait voir à la fois comment augmente le nombre d'éléments d'un environnement générateurs d'informations relationnelles traçables (le thème des objets connectés / capteurs), et comment se transforment les comportements des individus (humains et non humains) qui sont de plus en plus massivement émetteurs et récepteurs d'informations traçables. Les sciences sociales nous l'ont appris : quand on agit, on interagit. Et l'analyse économique nous a appris à graduer les interactions en fonction des informations dont disposent les acteurs pour agir, et en fonction de la symétrie ou de l'asymétrie informationnelles des relations d'interaction et d'échange. Or le puissant paradoxe qui est inhérent au développement des outils et des plates-formes numériques réside dans la construction et dans l'exploitation d'une asymétrie à laquelle les usagers et les consommateurs consentent : les outils sont relationnels, mais, sans l'asymétrie qu'ils exploitent, ils perdent leur efficacité. Cette asymétrie est invisible : les usagers ne savent que bien peu de choses des plates-formes, des algorithmes et de toute l'économie numérique

qui transforme en valeur les innombrables données personnelles recueillies sans cesse sur eux, au point que l'on a pu conceptualiser cette production de données captables comme un travail « bénévole à but lucratif » (Dujarier, 2016), fourni par les usagers des objets connectés. De multiples controverses et des demandes croissantes de régulation cherchent à rétablir la frontière entre ce qui appartient à l'individu et ce qu'il consent à laisser observer de lui. Mais le jeu de l'innovation déplace sans cesse ce pouvoir de l'asymétrie. L'économie numérique, incarnation d'une nouvelle révolution industrielle, représente aussi une concentration de l'expertise et de la créativité qui donne un pouvoir économique et technique intimidant à une jeune génération d'entrepreneurs, d'inventeurs, d'ingénieurs et d'informaticiens.

Au regard de cette émergence, quels gains peut-on attendre de cette nouvelle forme, graduable, de souveraineté des individus ? L'accès à des informations auparavant introuvables ou inexistantes, les facilitations marchandes, la sociabilité démultipliée, les innovations attachées aux externalités de réseau ? Et quels en sont les risques ? Trop de publicité, trop de profilage, trop de prédiction segmentatrice enfermant les consommateurs dans un espace réduit de choix, trop d'avantages cumulatifs pour certains biens ou services, laissant une multitude d'autres biens et services dans un bruit informationnel indistinct ? Quelles sont les résistances, les controverses et les révoltes ? La réémergence des débats sur la protection de la vie privée, la définition de biens communs inappropriables, les nouveaux horizons politiques du piratage informatique ? Les services secrets disposent de leviers extraordinaires avec l'exploitation des données de masse, et jouissent d'un monopole légal dans l'exploitation de l'asymétrie d'information, autrement dit d'une nouvelle forme de monopole de la violence légitime, pour reprendre la définition donnée par Max Weber de l'État et de son usage légal de la force. Admettra-t-on alors que les entreprises et les organisations qui peuvent prélever, gérer et revendre des données personnelles bénéficient d'une asymétrie d'information légitime, et contractualisable ?

Finalement, cinq évolutions se superposent désormais :

- les innovations scientifiques et techniques ;
- la multiplication des sources génératrices de données et leur connectivité ;
- l'expansion du marché des outils et des usages d'exploitation des données de masse ;
- la mise en controverse éthique et politique, et la mise en régulation juridique et législative, des déséquilibres informationnels exploités ;
- la formation de nouveaux marchés de services transformant la *privacy*, la vie privée, en utilité ou « désutilité » graduable et négociable.

***

Le 2 juin 2014 s'est tenu au Collège de France le colloque intitulé « *Big data*, entreprises et sciences sociales. Usages et partages des données numériques de masse », dont Simon Paye et moi-même avons pris l'initiative. On ne compte plus les publications sur le caractère « révolutionnaire » du phénomène des *big data*[2]dans le monde de la recherche. Avec un ensemble de collègues dont les contributions constituent la matière de ce livre, nous avons voulu explorer

2. Citons ainsi, parmi les travaux examinant l'impact de cette révolution des données sur la recherche en sciences sociales, King (2009, 2011), Lazer *et al.* (2009), Bail (2013), National Research Council (2013), Ruppert, Law et Savage (2013), Einav et Levin (2014), Kitchin (2014), Wyly (2014), Venturini *et al.* (2014).

les positions et les défis des sciences sociales face à la prolifération des traces numériques et de leur marchandisation. Il est par exemple annoncé que les données numériques de masse ouvriront la voie à une science sans théorie : plus besoin de s'embarrasser d'hypothèses et de références scientifiques si les données, par leur incroyable richesse, permettent de suivre en temps réel l'évolution des comportements, à l'instar d'une sorte de météorologie du social. De même, les acteurs du marketing, des assurances ou du conseil en organisation promettent l'avènement d'une nouvelle ère dans laquelle l'entreprise devenue intelligente peut prendre des décisions fondées sur une quasi-omniscience de ses dynamiques internes et de son environnement. Pour ceux qui tiennent ces discours, les traces numériques des individus connectés ne sont rien de moins qu'un nouvel Eldorado informationnel.

L'ouvrage examine, dans une première partie, les usages des données numériques et les logiques qui les sous-tendent, qu'elles soient marchandes, scientifiques, policières ou communautaires. Puis les pratiques concrètes et les opérations au quotidien sont analysées dans une deuxième partie, à partir d'enquêtes de terrain localisées. Les contributions de cette partie explorent en détail les comportements et les modèles émergents, tout en menant une réflexion parallèle sur les enjeux sociaux, éthiques, politiques auxquels ils donnent lieu. Une troisième partie est consacrée à l'utilisation des données numériques dans le cadre de la recherche en sciences humaines et sociales. L'enjeu est de clarifier, à partir de cas concrets, les perspectives que ces nouvelles sources de données ouvrent à la sociologie, à l'économie, aux sciences humaines, sans négliger les problèmes pratiques d'accès, de stockage ou d'interprétation qu'elles soulèvent. Enfin, une postface interroge la place du travail humain dans l'industrie du numérique.

La brève présentation des chapitres qui suit permettra de faire apparaître la triple ambition de cet ouvrage collectif : 1) donner à comprendre le rôle de l'asymétrie des relations utilisateurs / plates-formes dans la captation, l'exploitation et la marchandisation des traces numériques ; 2) rassembler des connaissances empiriques nouvelles sur la genèse et le développement de marchés liés à l'exploitation des données numériques, en faisant apparaître les différents intermédiaires qui s'insèrent dans les chaînes de valeur ; 3) offrir une synthèse de nos réflexions sur les enjeux cognitifs, épistémologiques et méthodologiques des usages de données numériques de masse en sciences sociales.

Frank Cochoy et Jean-Sébastien Vayre proposent une généalogie des relations d'échange qui sont au fondement de la production et de l'utilisation des données massives. Sur un marché opèrent des entreprises et des consommateurs. Ce marché est lui-même la somme des relations d'échange que les entreprises entretiennent avec d'autres entreprises, et des relations que les entreprises nouent avec les consommateurs individuels. Dans tous les cas, la vente de biens et de services fait appel à des dispositifs destinés à faire connaître les produits, à les valoriser, à persuader les acheteurs de les adopter sans en avoir fait l'expérience préalable, et à les convaincre de devenir fidèles, à l'aide d'arguments visant à soutenir une réputation de qualité, pour un rapport donné entre prix et caractéristiques du bien. Chaque concurrent agit de même. Le marketing et ses analyses des caractéristiques des consommateurs, de leurs besoins et de leurs préférences, la publicité, les techniques de fidélisation de la clientèle sont autant de dispositifs générateurs de données à prélever, à exploiter et à faire remonter vers le producteur. Les dispositifs numériques de prélèvement

massif de données personnelles, de traitement et d'exploitation marchande de celles-ci ne sont qu'une extension de ces dispositifs, selon Cochoy et Vayre. Les auteurs détaillent ce traçage généalogique en insistant notamment sur les savoirs mobilisés et sur les techniques mises en œuvre pour transformer l'individu consommateur ou usager (dans le cas des services publics) en grandeur statistique et pour identifier, grâce au calcul probabiliste, les comportements prévisibles, les préférences influençables, les demandes à satisfaire et les innovations profitables. Avec les techniques numériques, les individus ne sont plus simplement des représentants de classes statistiques issues de patients travaux économétriques de classification et de segmentation, ils sont désormais identifiés directement, par leurs caractéristiques personnelles, par leurs comportements et leurs trajectoires de consommation, par leurs préférences et par leurs relations. Ils sont traçables, visualisables et susceptibles d'être individuellement influencés. Qui donc traite toutes ces données pour alimenter tous les dispositifs d'action sur le consommateur ? Une armée d'ingénieurs et de statisticiens ? Non, des machines programmées pour se rendre autonomes et apprendre à travailler seules. Les discours sur les innovations sont incessants, mais l'efficacité de ces dernières est plus incertaine que ne le promettent les multiples sociétés de conseil et les fournisseurs de services numériques qui enjoignent aux entreprises de ne pas manquer le virage des *big data*. En replaçant les pratiques de marchandisation des données numériques dans le cadre d'un « marché des dispositifs marchands », les auteurs se donnent les moyens d'interroger à nouveaux frais la performativité de ces promesses.

Par le titre de sa contribution, Bernard Harcourt suggère d'emblée une parenté théorique avec l'œuvre de Michel Foucault. Qui détient un savoir détient un pouvoir. Le savoir produit à partir des données massives est traité, exploité et échangé dans des conditions, et pour des usages, dont chaque personne privée n'a au mieux qu'une infime idée. L'économie biface des plates-formes repose, comme on le sait, d'un côté, sur la mise à disposition d'un service gratuit ou très peu coûteux pour le consommateur, et, de l'autre, sur le consentement de ce consommateur à abandonner tout contrôle sur les informations que produit son utilisation du service. Mais cet échange n'est que la pointe d'un iceberg d'opérations et d'échanges dont le volume et la variété ont très rapidement et considérablement augmenté. Pour les connaître, le journalisme d'investigation, l'expertise citoyenne, le travail scientifique militant et l'enquête historienne offrent leurs solutions et leurs lots de révélations, comme le montre Harcourt au début de sa contribution. Mais, objectera-t-on, les savoirs produits à partir des données individuelles ne sont pas intrinsèquement condamnables, puisqu'ils peuvent servir le bien commun. C'est l'argument qui soutient l'édifice entier des différentes sciences impliquées dans la connaissance des individus, de leur santé, de leurs comportements et de leurs préférences. La mise en place de normes procédurales et de garanties éthiques voudrait permettre de dépasser les difficultés liées à l'asymétrie entre l'expertise du scientifique et le défaut d'expertise du profane. Et la connaissance qui sera issue de l'exploitation de ces sources numériques ne peut devenir un bien public que si sa production est soumise au libre examen de ses méthodes, de ses résultats et de ses usages possibles, et si son utilisation est non rivale et non exclusive.

L'enquête, nous montre Harcourt, doit se déployer sur les deux versants : qui agit et pour quel usage ? Une fois fixées les équations de base de la science et de l'industrie des données (quantités produites, quantités à stocker, catégories d'opérateurs, capacités tech-

nologiques de traitement), une économie politique des données doit différencier les types d'informations recueillies et traitées, en fonction des types d'entités émettrices de données – les humains, les machines, les entreprises, les objets, les administrations – et de la nature des relations qu'entretiennent ces entités. Au plus loin de ce que peut être, à l'ère numérique, la production de biens publics de connaissance, et au plus loin de l'alignement que celle-ci suppose entre les rationalités économique, politique et démocratique, Harcourt montre comment la connaissance extraite des données est traitée comme un bien marchand et une ressource de surveillance sécuritaire. Le marché développe ses chaînes d'intermédiation, au long desquelles les données personnelles deviennent des biens de transaction ordinaires, et les caractéristiques des individus des variables ordinaires de tarification différentielle, alors qu'elles peuvent concerner les aspects les plus intimes et les plus vulnérables de la personne. Comme le montrent aussi Cochoy et Vayre, le marketing dispose ainsi d'instruments sans précédent d'analyse des comportements de consommation. Mais où s'arrête le marketing et où commence le contrôle exercé sur le consommateur ? Le rapprochement avec la surveillance d'État pratiquée par la *National Security Agency* (l'Agence nationale de sécurité américaine) et son programme PRISM souligne que les mêmes sources de données et les mêmes techniques d'analyse sont au service de pratiques de profilage et de probabilités prédictives, dont les buts sont différents, mais les rationalités semblables. Le défi démocratique que représente la régulation d'une économie politique des données aussi puissante et aussi aisément nourrie d'innovations incessantes est réellement considérable.

Guillaume Tiffon se demande comment qualifier l'activité des internautes dès lors que les traces informationnelles des comportements de ceux-ci ont pour les entreprises et leurs plates-formes une valeur directement capitalisable et exploitable. L'internaute n'est-il que le nouvel avatar du client mis à contribution de multiples manières pour interagir avec des machines qui se sont substituées à des personnels ? Avant de répondre, il faut d'abord identifier les diverses modalités de ce que peut être la « mise au travail du client ». Se fondant sur un ensemble d'enquêtes de terrain qu'il a réalisées, Tiffon détaille les diverses contributions qui sont apportées par les clients, sans être contractualisées ou monétarisées, pour accroître l'efficacité productive, dans la grande distribution, dans les centres d'appel, dans la restauration rapide ou dans les cabinets de kinésithérapie. Les multiples formes de *self-service* et d'interaction avec des automates (caisses, distributeurs, bornes d'enregistrement, etc.) en sont les exemples les plus familiers, mais ce ne sont pas les seuls. Ces contributions sont bien un travail qui demande au consommateur du temps et des compétences pour se rendre complémentaire des machines ou pour gérer des interactions dans des relations de service. Tiffon conclut que, parmi les définitions usuelles du travail, y compris celle qu'avait Marx de la création de valeur par extorsion de la plus-value, aucune ne semble s'appliquer à l'activité de l'internaute. Ce n'est pas le Marx de la théorie de la valeur-travail qui aide à penser la position de l'internaute devant un écran dont l'envers est un prédateur d'informations, mais plutôt le Marx contempteur de la consommation, et de cet autre type d'aliénation qu'est la manipulation des besoins individuels par le marché – et plus encore Foucault, mis en avant par Harcourt, pour ses analyses sur la mise sous surveillance des individus par des dispositifs de contrôle et de pouvoir auxquels ils sont, au mieux, invités à consentir en grande, ou même en complète, méconnaissance de cause.

Éric Dagiral et Sylvain Parasie examinent l'émergence de la science des données, son épistémologie si particulière de l'induction, et les comportements des praticiens actuels. En enquêtant auprès de *data scientists*, en revisitant la littérature académique et en analysant la conception et les évolutions d'un des logiciels favoris des praticiens, les deux auteurs montrent que les *data sciences* sont au point de convergence d'une série de transformations. En statistique, dès les années 1960, l'intérêt se tourne vers une démarche exploratoire. L'idéal de précision et de rigueur qui fait de la statistique un simple champ d'application des mathématiques conduit à sacrifier d'autres heuristiques, comme l'exploration et l'induction. Or les outils informatiques ont précisément pour vertu de permettre de manipuler rapidement de grandes quantités de données, pour y détecter des relations et des patterns, dont la signification n'est pas tirée d'un modèle hypothético-déductif, mais approchée et interprétée par tâtonnements. Le développement de l'informatique met de multiples logiciels statistiques entre les mains de non-statisticiens. La collecte, la gestion et le traitement de données, d'abord considérés comme des opérations techniques d'application, deviennent innovants : pratiqués par des informaticiens et des firmes informatiques, ils font naître des outils d'analyse qui ne dérivent pas simplement de recherches académiques. Enfin, quand, au début des années 2000, la science des données se constitue, c'est par une levée des frontières entre statistique et informatique, et entre chercheurs universitaires et chercheurs des entreprises. Stanford attire par exemple des économistes qui veulent accéder à des données d'une variété et d'une ampleur inédites, en collaborant avec des ingénieurs et des statisticiens. Dans la recherche génétique et biologique, les collaborations entre l'université et les entreprises se sont considérablement développées et sont un levier décisif pour l'innovation. La science des données s'émancipe quand une diversité grandissante d'acteurs contribuent à ses avancées. C'est ce que montre l'étude des différents apports à l'évolution du logiciel très populaire R, ou encore l'accueil réservé à la *data science* dans des milieux très divers, mais tous directement intéressés par l'exploitation de grandes quantités d'informations – journalistes, *marketers*, chercheurs en biomédecine, etc.

Quand les parties prenantes se multiplient, elles déploient des savoirs et des technologies qui n'obéissent plus à un modèle dominant de prélèvement et d'exploitation des informations individuelles. Un monde de transactions émerge. Les données peuvent être prélevées sur les individus à leur insu. Mais les données émises par chaque individu constituent aussi un bien ou un service qui pourra lui être vendu pour lui permettre de connaître et de contrôler son environnement, sa santé, sa sécurité, ses déplacements, ses relations, ses transactions, ses interactions de sociabilité : c'est le *quantified self*. Or la qualité de ce service marchand de quantification individualisée dépend de l'exploitation à grande échelle des données recueillies, qu'il s'agisse de coordonner les comportements, d'analyser les risques collectifs, d'identifier les mécanismes de contagion ou encore d'augmenter le pouvoir prédictif des connaissances accumulées. Un ensemble de techniques et d'algorithmes qui nous renseignent sur notre état de santé, sur les qualités de notre nourriture, sur les variations de notre état physique au travail et sur toute autre dimension de notre comportement augmentent la puissance des services marchands qui nous sont vendus, mais ils peuvent alimenter aussi les bases de données qui sont exploitées par la recherche publique à des fins d'intérêt général. Comment établir une distinction stable et efficace entre ce qui sera extrait de nos données

personnelles pour améliorer le bien public fondé sur la connaissance, et ce qui, à travers les multiples canaux de diffusion possibles des données, alimentera des usages commerciaux et servira des intérêts particuliers? L'individu émetteur de données personnelles est face à une asymétrie sans solution : comment peut-il concevoir tous les usages possibles de la quantité et de la diversité considérables de données qui le concernent ou l'incluent, afin de décider d'être non seulement un consommateur souverain, mais aussi un producteur souverain d'informations exploitables ?

Finalement, le contraste est grand entre l'histoire longue des hybridations mêlant la statistique et l'informatique, telle qu'elle est relatée par Dagiral et Parasie, d'une part, et la mise à l'agenda récente et spectaculaire de la question des données numériques de masse dans les organisations de recherche, les entreprises, les administrations et l'arène publique, d'autre part. Les attentes sont considérables, les injonctions à la modernisation sont lancées quotidiennement, les espoirs d'avancées scientifiques sont devenus omniprésents, mais les controverses et les critiques sont aussi très vives contre la transformation des individus en matrices d'informations exploitables sans limitations ni contrôles efficaces.

David Pontille et Didier Torny procèdent à une analyse généalogique et à un état des lieux de l'outil bibliométrique, devenu l'un des socles de l'évaluation de la qualité scientifique des chercheurs et des institutions, et l'un des instruments de pilotage de la recherche et de la gestion des recrutements et des carrières. Ils distinguent ainsi trois configurations qui ont situé la science elle-même à l'avant-garde de l'exploitation de données massives, et ils montrent que l'inventivité algorithmique et métrologique est affaire de jeux d'acteurs, de variété des sources de données et de stratégies concurrentielles. On connaît l'une des fonctions traditionnellement remplies par les revues scientifiques : celles-ci se substituent, pour une large part, aux universités et aux instituts de recherche pour évaluer, sélectionner et valider la production de recherche. La scientométrie fut mise au point par Eugene Garfield, qui s'appuyait sur les travaux théoriques de Derek J. de Solla Price. Elle a consisté à mesurer la valeur des revues et la valeur des articles à leur audience (à leurs citations), et à introduire ces informations dans les outils d'évaluation des individus, de hiérarchisation des revues et de classement des universités. L'analyse de Pontille et de Torny retrace les initiatives des firmes qui ont concurrencé l'entreprise pionnière de Garfield, qu'il s'agisse d'éditeurs comme Elsevier, ou du moteur de recherche Google Scholar et des indices bibliométriques qu'il produit. Parce que le but d'une recherche est d'être aussi accessible et visible que possible, et que la traçabilité de son appropriation est un progrès irrécusable, la production scientifique engendre un monde de données idéalement renseignées, idéalement individualisables, et idéalement exploitables. L'ingénierie bibliométrique a développé, à partir de ce matériau de qualité exceptionnelle, une cascade d'innovations pour étendre son emprise de mesure et d'évaluation à la plus large production scientifique possible, et pour proposer une grande variété d'indices et de systèmes de cotation de la valeur des supports de publication. Ici aussi, la situation apparaît de part en part asymétrique. Les revues n'existeraient pas sans la matière première des travaux scientifiques dont elles publient les résultats. Mais l'accès à ces revues est souvent payant, et les bouquets d'abonnement sont coûteux. L'utilisation systématique des données bibliométriques pour mesurer la productivité ou pour cerner le potentiel des individus peut se substituer au grain fin des procédures collégiales d'évaluation

qui font appel à une pluralité de sources d'appréciation. Les revues se multiplient, et sollicitent autant qu'elles absorbent la production d'une population mondialisée de chercheurs. Les algorithmes bibliométriques se diversifient pour rapporter la valeur d'une production scientifique à d'autres critères d'audience et d'impact que sa citation dans le seul périmètre des revues. Progressivement, le jeu de la science fait face à des dilemmes radicaux : faut-il consolider le pouvoir marchand des éditeurs jusqu'à transformer l'évaluation scientifique en une somme de technologies métrologiques et indicielles qui alimentent un marché de services destinés aux chercheurs et à leurs institutions pour augmenter leur productivité et leur impact ? Ou faut-il libérer complètement l'accès aux productions scientifiques à travers des archives ouvertes, qui suspendent le pouvoir sélectif et le travail éditorial des pairs, quitte à pousser chaque scientifique à devenir le gestionnaire de sa propre visibilité et à activer à cette fin de multiples sources d'impact, d'audience, d'attention, d'usage ? Dans ce dernier cas, rechercher et mesurer l'attention des publics les plus variés revient-il à leur reconnaître une expertise profane légitime, qui complète l'expertise professionnelle des pairs, voire s'y substitue ?

Dans leur contribution, Jérôme Denis et Samuel Goëta étudient le phénomène de l'ouverture des données publiques (*open data*), contemporain de celui données massives, et précisent les modalités de son expansion. L'ouverture des données est le foyer d'une mobilisation qui oppose directement l'accès public à l'exploitation industrielle et commerciale des informations, données et connaissances en jeu. En théorie, un individu ou une organisation producteurs ou émetteurs de données qui ont une valeur pour autrui peut soit s'en réserver l'utilisation, soit autoriser celle-ci en s'assurant du contrôle des usages qui seront faits des données, soit en ouvrir complètement l'utilisation si ces données sont considérées comme des biens publics, ni rivaux ni exclusifs. Dans tous les autres cas, quand l'émetteur des données ne peut exercer qu'un contrôle partiel, sporadique, ou même aucun contrôle du tout, sur l'usage et sur la transformation des informations qu'il a émises, le partenaire de l'échange (plate-forme, site web, moteur de recherche) tirera un avantage de la collecte des informations auprès de chaque individu. Cet avantage est certes infinitésimal, si chaque cas est considéré isolément, mais il devient potentiellement considérable s'il agrège ces informations avec celles obtenues auprès de tous les autres utilisateurs du service. Ce levier essentiel de l'exploitation à grande échelle d'une asymétrie informationnelle est un cas classique de fonctionnement imparfait des marchés : un acteur en sait davantage que son partenaire dans l'échange sur les profits qu'il peut tirer de cet avantage, alors même que les deux parties sont en relation d'échange et que l'échange devrait être transparent pour être équilibré et efficient. Le phénomène de l'Open Data ne se présente pas comme une solution à de telles défaillances, puisqu'il concerne d'abord toutes les données qui ont vocation à devenir des biens publics d'information et de connaissance, que ces données soient produites par des acteurs publics (administrations publiques ou autres), par des acteurs dont les activités sont financées par des fonds publics (chercheurs scientifiques notamment) ou par des producteurs de connaissances et d'outils techniques désireux d'en rendre l'utilisation totalement libre (créateurs du Web 2.0, militants des logiciels

libres). Pourtant, le motif de l'asymétrie informationnelle réapparaît quand il s'agit de définir les caractéristiques désirables de ces données ouvertes.

C'est assurément toute une nouvelle organisation de l'économie de la connaissance et de l'innovation qui veut imposer ses exigences : exigences de démocratisation, de transparence des opérations de production et de traitement des informations, mais aussi exigences de libre exploitation d'une matière aussi proche que possible de son état brut, afin que son traitement, qui sera générateur de valeur ajoutée, puisse préserver un lien traçable avec la source. Mais, comme le montrent Denis et Goëta, les données présentées comme brutes sont en réalité des données « brutifiées ». L'opération de « brutification » consiste à transformer des données de travail originales, construites et formatées selon des règles et des besoins internes, en données génériques susceptibles de traitements statistiques. Mais ne revient-elle pas à rétablir un filtre permettant aux administrations de conserver une marge de discrétion ? Si « brutification » équivaut à standardisation, la démocratisation de l'accès aux données conduit-elle si heureusement qu'on le dit à la diversification des sources et à la démocratisation des expertises qui peuvent s'en saisir ? Si, selon l'argument bien connu, une concurrence plus parfaite procure des gains d'efficience, le rétablissement d'un équilibre informationnel peut permettre d'améliorer le fonctionnement des marchés de biens et de services, de multiplier la production de connaissances, et de décentraliser le pouvoir d'innovation. Il doit aussi rendre les acteurs publics plus directement et plus continûment responsables de leurs actions, ce qui rapproche alors l'ouverture des données d'un dispositif de contrôle propre au principe de l'audit, dont l'extension est pourtant controversée. Enfin, il participe de l'émergence d'une *crowd science* ou science citoyenne, qui, selon les cas, sera complémentaire ou rivale de la science professionnelle. Cette dernière peut au demeurant trouver elle-même de nouveaux leviers d'innovation dans une circulation ouverte des données publiques, tout en ayant à chercher son propre équilibre entre la culture de l'accès et la protection du monopole temporaire de l'innovateur.

La sociologie est, dès sa fondation à la fin du XIX^e^ siècle, une science des relations et des interactions autant qu'une science de la détermination des actions et des comportements des individus et des groupes. La sociologie des réseaux a connu un essor considérable à partir des années 1970 aux États-Unis. Elle pouvait fonder son développement sur l'une des inspirations originaires de la sociologie américaine, celle qui, comme l'ont souligné successivement John Dewey, George Herbert Mead, Herbert Blumer, Erving Goffman, Howard Becker, Harrison White et Andrew Abbott, fait de « l'acte social », ou de l'action immergée dans un réseau sans cesse changeant d'interactions, le fondement de ce qu'est une société (Menger, 2013). La sociologie des réseaux a formalisé et mathématisé cette intuition originaire. L'invention du Web 2.0 pouvait livrer à cette sociologie relationnelle une importante matière première et des situations de quasi-expérimentation, fondées sur les ressources illimitées d'interactivité, mais aussi sur une immersion des individus dans un monde de relations entre des entités de types multiples, dont les humains ne sont qu'une catégorie. Une technologie conforme à une ontologie whiteheadienne du monde social (Menger, 2016) semble être apparue.

À partir des travaux du laboratoire Orange Labs dont il est membre, Jean-Samuel Beuscart procède à un bilan des usages scientifiques des données du Web, et détaille les avancées et les limites du déploiement de cette sociologie dans les recherches sur les usages des nouveaux

médias. Le Web paraît abolir toutes les contraintes sur la formation et le développement des relations entre toutes les entités qui peuplent la Toile, et favoriser un monde horizontal, sans hiérarchies établies *a priori*. Mais les inégalités d'attention, de visibilité et de réputation que reçoit chacun existent bel et bien et nous rappellent que le Web obéit à des principes de gravitation relationnelle : ces principes sont bien différents de ceux d'une démocratie horizontale de liberté d'accès et d'échange. Les comportements des internautes participatifs sont en effet distribués très asymétriquement : la connectivité, tout comme la réputation au sein de la communauté des participants, obéissent à une loi de puissance bien connue qui veut que beaucoup contribuent peu, et peu contribuent énormément. Beuscart rappelle aussi que la matière première des liens, des traces, des usages des sites participatifs, qui est disponible pour l'analyse sociologique, n'est pas un or brut, mais un ensemble d'actes et d'informations déjà modelés en fonction de l'administration des contacts par les sites du Web 2.0. Ni objectivité ni exhaustivité des données à extraire : la mise en garde est salutaire, et elle permet aux sociologues de se maintenir en état de vigilance épistémologique tout en innovant dans le traitement de cette matière sociale considérable des actes relationnels. C'est l'équilibre entre contrôle empirique, expérimentation méthodologique et innovation théorique qui peut permettre d'avancer dans l'étude de multiples questions classiques de la science sociale, telles que les pratiques culturelles, le jugement de goût, la constitution de l'expertise profane, la construction des réputations ou encore l'influence des consommateurs sur le comportement des marchés de biens et de services infiniment différenciés. La recherche peut ainsi montrer comment les individus échangent des jugements et des évaluations, en contournant les hiérarchies anciennes qui réservaient l'émission d'opinions expertes à des professionnels du jugement et de la critique, mais elle montrera aussi comment ces nouvelles formes d'échanges plus horizontaux peuvent enclencher des dynamiques d'affiliation des jugements dont les marchés cherchent à influencer le cours.

Que valent ces nouvelles données ? Complètent-elles l'information recueillie par les procédés ordinaires d'enquête directe sur les comportements relationnels et sur les déterminants des préférences, des jugements et des choix, ou peuvent-elles prétendre s'y substituer en réorientant l'agenda des recherches vers de nouvelles classes de faits sociaux ? À la lecture de la contribution de Beuscart et de celles des autres auteurs du volume, il me semble que quatre réponses émergent. Premièrement, les données relationnelles proposent surtout une géométrie du social, puisqu'elles renseignent le plus souvent sur des comportements et sur des interactions sans permettre de caractériser finement les acteurs au-delà des traces et des informations qu'ils produisent. En ce sens, la carte des réseaux et des dynamiques de liens entre les individus ne saurait prétendre fournir une représentation directe de la société et de ses transformations. Deuxièmement, des phénomènes sociaux dont on avait l'intuition sont brusquement révélés et documentés avec une ampleur empirique sans précédent, notamment pour ce qui touche à la tension entre la puissance d'échange horizontal entre les individus et les hiérarchies qui se constituent de manière endogène à partir des inégalités d'usage, d'influence, de visibilité, de contribution, de réputation. Troisièmement, autant que la morphologie et la dynamique des réseaux et de leurs effets sur les comportements individuels, c'est une nouvelle matière de l'action sociale, la matière informationnelle, qui s'offre à l'analyse sociologique, avec une densité et une complexité inédites. Enfin, la sociologie, comme d'autres sciences

sociales, peut construire de multiples liens de collaboration avec les sciences informatiques pour innover, et pour s'appliquer à elle-même le principe fondateur de la sociologie relationnelle : le changement et la transformation sont la donnée primitive du système d'action, et c'est l'inertie qui constitue le problème à explorer et à résoudre.

Il ne faut sans doute pas négliger le phénomène de mode derrière l'engouement pour ces nouveaux gisements de connaissances promises. Mais le diagnostic suppose de ne pas basculer pour autant dans la mode de la critique négatrice (Cardon, 2015). Qu'en est-il, au juste, des innovations pour les sciences sociales ? Les protocoles de recherche se modifient quand les matériaux s'offrent en quantités illimitées. L'heuristique de l'analyse relationnelle et de la détection de formes gagne ses lettres de noblesse : de nouvelles corrélations, des configurations de relations, des tendances et des régularités peuvent être mises au jour, visualisées et interprétées. Mais la promesse d'une innovation de grande portée pour la science sociale est-elle tenue ? Une nouvelle physique du monde social est-elle en train de naître ? Pour l'instant, on n'en discerne pas les traits, du moins tant qu'on se contente d'assimiler la révolution des données à l'analyse des traces numériques des individus. En réalité, il faut, selon Ollion et Boelaert (2015), distinguer trois sources bien différentes de données : celles produites par les usages d'Internet ; celles, certainement plus décisives pour les sciences sociales, qui proviennent des actes de travail ainsi que de la production et de l'échange d'informations dans les entreprises, dans les organisations, dans les administrations ; et celles qui proviennent de la numérisation de sources documentaires prénumériques. Comment le travail du chercheur en sciences sociales est-il affecté par l'abondance et la variété de ces sources ? Les nouveaux impératifs sont cognitifs (l'acquisition de nouvelles compétences), heuristiques (la détection de patterns, et la tolérance au flottement inductif dans l'exploration des données) et méthodologiques (l'importation de méthodes familières à d'autres disciplines). Ils sont aussi éthiques et juridiques, puisqu'il s'agit bien de protéger les données personnelles exploitées par la recherche tout en favorisant l'accès à ces ressources nouvelles. Enfin, comme le montre le foisonnement des recherches exploitant les traces numériques, la matière sociale des comportements et des relations est accessible à une variété croissante d'analystes, physiciens, informaticiens, mathématiciens, mais aussi experts en maniement des données, professionnels de l'information et journalistes. Comme le souligne Jean-Samuel Beuscart, si les individus laissent une masse de traces, celles-ci sont de faible épaisseur sociale. L'ivresse « quantophrénique » ne doit alors ni masquer l'insuffisance de nouvelles propositions théoriques, ni faire oublier l'importance cruciale des données produites dans le contexte d'une recherche, quelle que soit la méthode retenue : observation, entretiens, questionnaires, etc. (Venturini *et al.*, 2014).

Dominique Boullier s'inscrit différemment dans ce débat puisqu'il se donne pour ambition d'identifier trois générations de sciences sociales, en se fondant sur un ensemble de variables analytiques dont le tableau est présenté à la fin de son chapitre. Si la science sociale est une science probabiliste, il est tentant de dessiner des configurations dans lesquelles le principe probabiliste dépend, en premier lieu, de l'état des techniques de collecte d'information employées pour caractériser la position des individus dans une totalité sociale, en deuxième lieu, des objectifs qu'ont les acteurs intéressés à la production du savoir social, et, en troisième lieu, des méthodes et des outils de traitement et de

représentation des informations caractérisant le comportement des entités élémentaires dont on cherche à décrire et à expliquer le comportement. L'argument est ici que de la science sociale d'Émile Durkheim à celle de l'acteur-réseau de Michel Callon et Bruno Latour, une réduction d'échelle intervient progressivement. La relation de l'individu à la totalité sociale, les conditions d'équilibrage de la différenciation individualisante et de la solidarité interindividuelle, et les perturbations de celle-ci en présence de crises d'anomie sociale, étaient au cœur du système durkheimien. L'analyse des comportements cherche ensuite à devenir plus prédictive, avec la technologie des sondages et des mesures d'opinion, et explore des réalités plus mésosociologiques comme les phénomènes de diffusion et de contagion informationnelle, dans lesquels les caractéristiques des individus, mais aussi leur position dans des réseaux de relations interindividuelles et leur exposition variable aux sources d'information, sont prises en compte par les modèles. La focale se resserre encore quand une topologie relationnelle caractérise les entités agissantes à l'aide des flux d'informations qu'elles émettent, reçoivent, associent, transforment et propagent sans cesse. Ici, c'est Gabriel Tarde, le rival de Durkheim, qui, selon l'argument de Boullier, triomphe, avec sa théorie d'inspiration leibnizienne. Mais comment, dans une telle ontologie sociale de troisième ordre, loger tout ce qui est révélé par les autres contributions au volume : l'exploitation stratégique des asymétries informationnelles, les amplifications de visibilité et de réputation, les techniques de contrôle, de persuasion et d'influence, et la forte concentration du marché des technologies numériques et des algorithmes, avec la domination des GAFA (Google, Apple, Facebook, Amazon) ? Boullier se demande lui-même, à la fin de sa contribution, comment éviter de conclure trop banalement qu'un modèle du social se substitue simplement à un autre.

L'ouvrage se clôt sur l'entretien qu'a réalisé Simon Paye avec Émilien, un étudiant qui finance ses études en travaillant pour un intermédiaire à qui des firmes comme Google ou YouTube sous-traitent des lots de tâches. C'est un « travailleur de la donnée » (Bastard *et al.*, 2013). Les GAFA mettent volontiers en avant les vertus novatrices, créatives et anticonventionnelles de l'organisation du travail de leurs salariés[3]. Au-delà des slogans sur l'« uberisation » de l'emploi, qui ne sont que des commodités publicitaires ou demi-savantes, il faut décrire et évaluer les pratiques d'emploi et de travail qui, comme les formules « auto-entrepreneuriales » de prestation, déplacent plusieurs des bornes de la flexibilité contractuelle et statutaire. Ces formes de *self-employment* s'affranchissent de l'appareillage juridique construit pour distinguer le travail salarié du travail indépendant, et mettent en avant les souplesses de la rémunération à la tâche. Cette allocation du travail est courante dans les pays où opèrent les entreprises pour qui travaille l'interviewé. L'entretien est riche d'enseignements et permet de prolonger les résultats de recherches récentes (Scholz, 2013 ; Cardon et Casilli, 2015). Les plates-formes telle que Google et YouTube, qui donnent accès à une infinité de sites et de contenus, et qui en structurent la consultation par les classements hiérarchiques des pages, doivent en permanence

3. Les témoignages sur les pressions à la productivité et les techniques managériales mises en œuvre n'ont pas moins de relief. Voir l'enquête de Jodi Kantor et David Streitfeld sur les pratiques managériales d'Amazon, parue dans le *New York Times*, « Inside Amazon: Wrestling big ideas in a bruising workplace », 15 août 2015.

contrôler et évaluer les sites référencés, la qualité des contenus proposés, la conformité de ceux-ci avec des dispositions législatives et éthiques (le contrôle des sites pornographiques est une des tâches fastidieuses qu'évoque Émilien), l'efficacité des requêtes qu'il est possible d'y lancer, ou encore la qualité des services proposés par la plate-forme, comme la valeur des traductions réalisées par Google Translate. Des « modérateurs » sont embauchés pour réaliser ce travail d'évaluation, de contrôle, de surveillance, de nettoyage. Ils forment une armée de prestataires dont Émilien, l'un d'entre eux, décrit le travail de façon très suggestive.

Le travail semble pouvoir être exercé hors de toute subordination : Émilien travaille où il veut, segmente ses tâches comme il l'entend, pourvu que chaque unité de travail accompli puisse correspondre à un lot minimal de tâches déterminées, ajuste son niveau d'effort en fonction de ses autres contraintes et du temps à donner à ses études, choisit ses horaires et son rythme de travail, peut moduler son niveau de concentration. Son bilan est simple :

> C'est un jeu un peu. Ça demande pas une concentration extrême. […] J'ai peut-être moyen de faire le double horaire si j'arrive à mettre en place ce qu'il faut. Là ça sera juste parfait. […] Déjà rien que là, je gagne plus que ma mère, alors que je fous rien et qu'elle se casse le cul, je me dis que c'est magnifique. […] Elle est agent immobilière. Avec un deuxième compte je vais être payé 3200 balles pour 30 heures de taf par semaine, où je veux, quand je veux, comme je veux. C'est déjà mieux que 90 % des gens salariés dans le pays quoi.

Mais l'entretien révèle sous une lumière crue les conditions de cette apparente « insubordination ». La prestation de services réalisée par Émilien peut être interrompue à tout moment par l'entreprise cliente, en fonction des performances attendues et observées. Le travail d'Émilien est organisé et distribué par des algorithmes, et fait l'objet d'un contrôle précis : ses faits et gestes sont chronométrés, ce qu'il produit est régulièrement évalué. Les écarts par rapport à la performance attendue, s'ils se répètent, conduisent à la fin de la relation commerciale. Un jeu tactique s'installe, qui rétablit une partie de l'équilibre bilatéral de l'échange commercial : Émilien sait développer des routines pour moduler son effort, parce qu'il est certain de sa haute productivité dans ce travail.

C'est en réalité à une mise en abyme du travail numérique que conduit cet entretien : Émilien observe et corrige ce que font les algorithmes des plates-formes tout en travaillant pour une entreprise qui emploie des algorithmes pour distribuer, quantifier, gérer et évaluer le travail demandé à ses prestataires. Et ceux-ci, s'ils sont assez agiles, peuvent exploiter les marges de manœuvre du dispositif algorithmique. Le paradoxe qui émerge est éloquent : on fait intervenir l'humain pour corriger les imperfections des algorithmes, tout en mobilisant des techniques de captation et d'exploitation des traces numériques pour contrôler l'activité des travailleurs de la donnée.

Références

Abiteboul S. (2012), *Sciences des données : de la logique du premier ordre à la Toile*, Paris, Fayard / Collège de France ; en ligne : Paris, Collège de France, http://books.openedition.org/cdf/506.

Autor D. (2013), « The "task approach" to labor markets: An overview », *Journal for Labour Market Research*, vol. 46, n° 3, p. 185-99.

Autor D. et Handel M. (2013), « Putting tasks to the test: Human capital, job tasks and wages », *Journal of Labor Economics*, vol. 31, n° 2, S59-S96.

Bail C. (2014), « The cultural environment: Measuring culture with big data », *Theory and Society*, vol. 43, n° 3, p. 465-482.

Bastard I., Cardon D., Fouetillou G., Prieur C. et Raux S. (2013), « Travail et travailleurs de la donnée », *Internetactu.net*, http://www.internetactu.net/2013/12/13/travail-et-travailleurs-de-la-donnee/ (dernière consultation le 3 avril 2017).

Bok D. (2003), *Universities in the Marketplace. The Commercialization of Higher Education*, Princeton, Princeton University Press.

Brynjolfsson E. et McAfee A. (2014), *The Second Age Machine*, New York, Norton.

Cardon D. et Casilli A. (2015), *Qu'est-ce que le Digital Labor*, Paris, INA Éditions.

Cardon D. (2015), *À quoi rêvent les algorithmes : nos vies à l'heure des big data*, Paris, Seuil.

Deming D. (2015), « The growing importance of social skills in the labor market », Chicago, *NBER Working Paper,* n° 21473.

Einav L. et Levin J. (2014), « Economics in the age of big data », *Science*, vol. 346, n° 6210.

Kitchin R. (2014), « Big Data, new epistemologies and paradigm shifts », *Big Data and Society*, avril-juin, p. 1-12.

King G. (2009), « The changing evidence base of social science research », *in* King G., Schlozman K. et Nie N. (dir.), *The Future of Political Science: 100 Perspectives*, New York, Routledge, p. 91-93.

King G. (2011), « Ensuring the data rich future of the social sciences », *Science*, vol. 331, n° 11, p. 719-21.

Lazer D., Pentland A., Adamic L., Aral S., Barabasi A.-B., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M., Jebara T., King G., Macy M., Roy D. et Van Alstyne M. (2009), « Computational social science », *Science*, vol. 323, n° 5915, p. 721-723.

Mayer-Schönberger V. et Cukier K. (2013), *Big data. A revolution that will transform how we live, work, and think*, New York, Houghton Mifflin.

McSherry C. (2001), *Who Owns Academic Work? Battling for Control of Intellectual Property*, Cambridge (Mass.), Harvard University Press.

Menger P.-M. (2013), « La dramaturgie sociale du travail. Une conception interactionniste de la stratification », *in* Perrenoud M. (dir.), *Les Mondes pluriels de Howard S. Becker*, Paris, La Découverte, p. 206-237.

Menger P.-M. (2016), « Temporalité, action et interaction », *in* Demazière D. et Jouvenet M. (dir.), *Andrew Abbott et l'héritage de l'école de Chicago*, Paris, Éditions de l'EHESS, p. 145-170.

National Research Council (2013), *Frontiers in Massive Data Analysis*, Washington, National Academies Press.

Ollion E. et Boelaert J. (2015), « Au-delà des *big data*. Les sciences sociales et la multiplication des données numériques », *Sociologie*, vol. 6, n° 3, p. 295-310 (en ligne : https://sociologie.revues.org/2613).

Popp Berman E. (2012), *Creating the Market University. How Academic Science Became an Economic Engine*, Princeton, Princeton University Press.

Ruppert E., Law J. et Savage M. (dir.) (2013), « Digital devices: The social life of methods », *Theory, Culture and Society*, numéro spécial, vol. 30, n° 4.

Scholz T. (dir.) (2013), *Digital Labor. The Internet as Playground and Factory*, New York, Routledge.

Slaughter S. et Leslie L., *Academic Capitalism. Politics, Policies and the Entrepreneurial University*, Baltimore, The Johns Hopkins University Press, 1997.

Venturini T., Cardon D. et Cointet J.-P. (dir.) (2014), *Méthodes digitales. Approches quali/quanti des données numériques*, *Réseaux*, n° 188 (en ligne : https://www.cairn.info/revue-reseaux-2014-6.htm).

Wyly E. (2014), « The new quantitative revolution », *Dialogues in Human Geography*, vol. 4, n° 1, p. 26-38.

I. Cheminements des *big data* : technologies, marchés, échanges

Les *big data* à l'assaut du marché des dispositifs marchands : une mise en perspective historique

Franck Cochoy

Professeur de sociologie à l'université Toulouse Jean-Jaurès, professeur invité à l'université de Göteborg (Suède) et membre du Laboratoire interdisciplinaire de solidarités, sociétés, territoires (LISST, UMR CNRS 50193)

Jean-Sébastien Vayre

Docteur en sociologie, membre du Centre d'étude et de recherche travail, organisation, pouvoir (CERTOP, UMR CNRS 5044) et de l'équipe Travail, relation, action, marché, et espace public (TRAME)

Les gestionnaires connaissent bien l'opposition entre marché « B to B » (*Business to Business*) et marché « B to C » (*Business to Consumer*). Joanne Yates (2006) a insisté avec justesse sur la focalisation excessive des historiens de l'innovation sur le B to C, en vertu de l'idée que les nouveautés techniques seraient tournées ou tirées par la demande, et sur l'oubli du B to B, c'est-à-dire sur l'importance des échanges interentreprises dans l'essor des innovations. Les *big data* ou « données massives » confirment l'excellence de cette perspective, puisque le développement des outils actuels d'intelligence des données massives est à l'évidence soutenu par un marché entre professionnels, rapprochant entreprises consommatrices de données et opérateurs prestataires de services. Mais l'auteur a peut-être commis une erreur en opposant les marchés B to B et B to C, alors que dans les faits ces deux formes de places marchandes sont étroitement imbriquées, et que la dynamique de cette imbrication est précisément ce qui permet selon nous de rendre compte de l'essor et de la généralisation des innovations en question. À l'articulation des marchés B to B et B to C intervient en effet une forme moins connue, et pourtant cruciale, que nous proposons de nommer « marché des dispositifs marchands », qui consiste à vendre aux entreprises des dispositifs de vente de produits au grand public (vitrines, sondages, plans marketing, programmes de fidélité, etc.). Le marché des dispositifs marchands recouvre donc la relation « (B to [B) to C] » : on vend à des organisations marchandes – (B to B) – des services ou des équipements de captation de la clientèle d'après l'argument que ces derniers permettront de vendre des produits aux consommateurs – [B to C] : la première vente repose entièrement sur la promesse pourtant non garantie de la seconde, de sorte que les premiers manipulés sur un tel marché sont les candidats à la manipulation.

Dans ce chapitre, nous voudrions replacer les *big data* dans le cadre de ce « marché des dispositifs marchands », dont ils ne sont qu'un avatar parmi d'autres. Cela devrait nous permettre non seulement d'en saisir la dynamique (comme tout produit vendu sur un tel marché, les *big data* doivent construire leur nécessité à grand renfort de promesses techniques et commerciales), d'en décrire les spécificités et l'efficacité (ces promesses sont particulières et renvoient à des éléments précis), mais aussi de les mettre à distance (la question de la performativité des promesses reste ouverte ; les conditions de leur pleine intelligibilité méritent examen).

1. Aux origines des *big data* : une mise en perspective historique

1.1. Big Market

Depuis Adam Smith, l'économie théorique s'est appliquée à décrire les marchés comme un espace unifié, régi par des lois analogues à celles de la nature. Cette discipline a tellement bien réussi dans cette entreprise qu'elle est parvenue à faire oublier l'origine et la portée doublement politiques de son projet : d'une part, l'invention du marché autorégulateur, de la poursuite de l'intérêt et de la libre entreprise fut au départ conçue comme une utopie de libération visant à sortir de l'Ancien Régime, en présentant le marché comme une autre façon d'obtenir l'ordre social, fondée sur un mécanisme naturel et décentralisé d'allocation des ressources, censé dispenser *ipso facto* du recours aux tutelles politiques ou religieuses jusque-là mobilisées pour dominer les passions (Hirschman, 1980) ; d'autre part, la promotion du marché comme nature a fondé le coup de force proprement politique des économistes, qui ont réussi grâce à leurs modèles à se faire reconnaître comme uniques représentants du marché, comme experts indispensables de la « chose économique », seuls capables d'approcher les lois censées gouverner les échanges et donc d'instruire les hommes sur la façon d'organiser et de conduire leurs affaires (Cochoy, 2015a).

De ce point de vue, le marché théorique peut être considéré comme le premier avatar des produits qui sont vendus sur le marché des dispositifs marchands, à ceci près que c'est un produit très particulier et ambivalent : d'un côté, c'est une solution universelle censée régler le cours des pratiques économiques de tous les acteurs, en tous lieux et en tous temps (le marché libre doit être institué partout et préservé toujours [Polanyi, 1983]) ; de l'autre, ce n'est en rien un outil dont l'on peut jouer pour régler des problèmes particuliers : l'efficacité globale dont se réclame le marché a pour prix l'impuissance locale généralisée qu'il promeut. Les acteurs des marchés financiers le répètent à l'envi : personne ne saurait « battre » le marché ; derrière cette conviction, on trouve l'axiomatique néoclassique, selon laquelle aucun acteur ne peut à lui seul influencer les conditions de l'échange (Martin, 2005).

1.2. Big Business

Cette façon de voir a longtemps conduit les acteurs économiques à vivre l'expérience marchande et les fluctuations économiques un peu à la façon dont les agriculteurs subissent les caprices de la météorologie, c'est-à-dire sur le registre d'une fatalité impérieuse, tantôt favorable, tantôt contraire, à laquelle on ne peut que se soumettre faute de pouvoir en affecter le cours. Mais Alfred Chandler (1988) a montré comment et combien la naissance des grandes entreprises à la fin du XIX^e^ siècle – l'avènement du Big Business comme premier préalable à l'essor des *big data* ! – a changé la donne : d'un côté, les mastodontes industriels capitalistes, handicapés et fragilisés par le niveau considérable de leurs coûts fixes, étaient plus sensibles que les petits agents économiques aux risques des aléas marchands ; d'un autre côté, leur grande taille leur donnait la capacité d'inverser le rapport de force entre marché et entreprise : grâce à l'invention de nouvelles fonctions, expertises et procédures de gestion, il devenait possible pour une

même entité d'avoir prise sur le cours du marché, par exemple en internalisant les circuits de distribution à l'échelle du territoire national, comme dans le cas des chaînes de magasins (Tedlow, 1990). Pour la première fois, la « main visible » des managers prenait la main… sur la « main invisible » du marché.

Ce basculement est à l'origine de ce que nous avons appelé ailleurs le « jeu de la main chaude » (Cochoy, 1999), et que nous proposons de revisiter grâce à la notion plus nouvelle de « marché des dispositifs marchands ». Le jeu de la main chaude a pour enjeu la maîtrise du marché en vertu d'une rhétorique très particulière, qui consiste chaque fois à vendre aux entreprises commerciales les moyens d'atteindre un rapport plus adéquat et plus direct avec leur demande en supprimant d'anciens médiateurs, mais bien sûr pour leur substituer une nouvelle médiation, qu'un autre prestataire de techniques de « *matching* marchand » viendra concurrencer à son tour, au nom de sa propre solution et de l'instauration d'un rapport prétendûment plus étroit encore avec la clientèle, et ainsi de suite. Selon ce schéma, les grossistes ont évincé les marchands ambulants censés tourner la médiation de l'épicier, à grand renfort d'économies d'échelle et de gestion intégrée, puis les fabricants ont pris l'ascendant sur les grossistes et les épiciers, grâce aux produits de marque nationale soutenus par des packagings orientés directement vers les consommateurs, etc. (Strasser, 1989). Parallèlement, cette prolifération de dispositifs et de spécialistes de médiation marchande de plus en plus sophistiqués et nombreux s'est accompagnée à partir de la fin du XIX^e^ siècle d'une dissociation, d'une autonomisation et d'une externalisation de l'expertise marchande, avec le développement de la publicité et des études de marché, et bien sûr avec la multiplication des diplômes et des personnels spécialisés dans l'administration de ces nouveaux savoirs et savoir-faire. Les études de marché et la publicité prétendent remplacer les représentants, vendeurs et autres acteurs de terrain comme moyen de connaissance et de maîtrise immédiates des marchés, grâce à l'instauration d'une communication directe auprès des clientèles (pour la publicité) et grâce à l'auscultation tout aussi directe des préférences des consommateurs (pour les études de marché).

Ces nouvelles médiations se bouclent – les études de marché alimentent la définition des campagnes publicitaires – en prenant appui sur trois éléments fondamentaux : la presse, le gouvernement et la science.

1.3. *Mass media*

La presse, d'abord écrite puis radiophonique et télévisuelle (et plus récemment « en ligne »), est au fondement de l'expertise marchande, dans la mesure où c'est non seulement la connaissance de l'audience qui fonde l'élaboration des contenus (articles et programmes), mais aussi la vente de cette même connaissance auprès des annonceurs qui co-financent[1] les entreprises

1. La presse est vendue auprès d'un double public : son public et les annonceurs, d'où la notion de marché biface (Rochet et Tirole, 2003). Même dans le cas où la presse est gratuite pour le public et « vendue » aux annonceurs la notion de marché biface garde toute sa pertinence, puisque la presse n'a de valeur pour les annonceurs qu'en vertu du public qu'elle attire.

de presse et qui sont avides pour cela d'obtenir les moyens d'identifier les canaux les plus fréquentés par leurs cibles et les plus adaptés à leurs goûts[2].

Pour illustrer l'apport de la presse au jeu de la main chaude et au fonctionnement du marché des dispositifs marchands, prenons l'exemple de la revue *Progressive Grocer*. Cette revue, bien que méconnue en France, a l'avantage de présenter toutes les caractéristiques du jeu et du marché en question. *Progressive Grocer* est un organe de presse lancé en 1922 à l'intention des petits épiciers, auxquels il entend fournir des informations sur leur secteur d'activité et présenter des solutions pour améliorer leurs affaires, dans un contexte de concurrence croissante des chaînes de magasins puis, dans les années 1930, des supermarchés. Le magazine est distribué gratuitement à ses lecteurs potentiels et financé par la publicité. Ainsi, il vend à ses annonceurs un espace censé les aider à vendre des dispositifs aux commerçants (B to B) selon l'argument que ces derniers pourront, grâce à eux, mieux et plus directement qu'avec les solutions antérieures (main chaude), vendre davantage de produits aux consommateurs (B to C) : réfrigérateurs électriques en lieu et place des anciens coffres à glace, présentoirs vitrés en remplacement des vieux comptoirs opaques, étagères en libre-service à la place de vendeurs intrusifs, etc. (dispositifs marchands). La revue articule donc jusqu'au tournis la logique du marché B to B to C, c'est-à-dire d'un système de ventes promues sur des promesses de vente. Elle rend publiques les solutions commerciales, les rassemble, fédère la concurrence selon une logique de co-opétition (Bruno, 2012), et fonctionne donc en cela comme un véritable salon professionnel de papier. Ce fonctionnement s'appuie sur un usage certes assez simple, mais néanmoins très réfléchi et astucieux des données marchandes, d'une part *via* la mobilisation d'annuaires professionnels pour construire un lectorat de 50 000 distributeurs dès son lancement et porté à 75 000 en 1930, un argument crucial à l'intention des annonceurs, d'autre part – et surtout *via* l'usage des données comptables – pour opérer le tour de passe-passe consistant pour la revue à exhiber sa propre efficacité comme miroir ou promesse de celle de ses clients potentiels : la revue publie ainsi des graphiques récurrents sur la taille et la répartition de son lectorat, ainsi que sur la progression de son chiffre d'affaires, selon l'argument que son succès fonctionnerait comme indice de celui de ses clients (Cochoy, 2014). Enfin et surtout, ce genre de logique se retrouve du côté des annonceurs eux-mêmes, qui promeuvent dans les pages de *Progressive Grocer* censées les aider à vendre leurs produits, des produits qui eux-mêmes reposent sur des promesses de vente.

1.4. *Big cabas*

Prenons l'exemple sans doute le plus emblématique du chariot de supermarché : celui-ci, apparu en 1936 (Grandclément, 2006) et largement promu dans les pages de *Progressive Grocer*, est au départ nécessairement commercialisé à l'intention d'un marché dominé de façon écrasante par de tous petits magasins indépendants (en 1930, on compte 250 000 com-

2. Pour plus de détails sur l'histoire de cette intrication et de cette dynamique, qui fonde toujours l'économie des *big data* contemporaines, voir la contribution de Dominique Boullier, chap. 8.

merces indépendants contre 57 000 magasins appartenant à des chaînes), réalisant de faibles volumes de vente auprès de clients habitués au service et à la livraison à domicile (la vente en libre-service intégral ne concerne que 13 % des magasins en 1939), soit une cible dont le « besoin » en paniers à roulettes est pour le moins incertain. La promotion des chariots passe alors par la production d'argumentaires tous azimuts, orientés autant vers les professionnels que vers les consommateurs, ou plutôt vers les professionnels en vertu des avantages supposés du produit pour les consommateurs. Aux professionnels *stricto sensu*, dont l'on connaît les préoccupations sanitaires (à une époque où la vente en vrac prédomine et où la vermine constitue une menace très sérieuse), le souci pour les coûts, l'attachement au service et le faible espace de vente, on signale que les chariots métalliques sont hygiéniques (« *no germ* » est le leitmotiv de l'American Wire Form Co.), solides, durables et inoxydables, qu'avant d'équiper les clients ils peuvent servir aux épiciers pour préparer les commandes, voire pour faire office de présentoirs, et surtout que les chariots sont « pliables » (« The Folding Carrier Co. » est le nom choisi pour son entreprise par l'inventeur du dispositif !), afin de ne pas encombrer exagérément un espace de vente très restreint. Aux consommateurs, et donc aux professionnels *via* les avantages supposés pour les consommateurs, on promet « des ventes accrues par consommateur » (American Wire Form Co., octobre 1936) ; « davantage de son argent [celui de la cliente] dans votre poche » (United Steel & Wire Co., février 1938), « l'accroissement du volume individuel en fournissant une capacité plus grande par consommateur avec un confort accru » (United Stee & Wire Co., juin 1938). Ces arguments portèrent tellement bien, la manipulation des épiciers par la promesse de manipulation des consommateurs fut si efficace, qu'en 1956 on comptait en moyenne près de 9 chariots par petit magasin, soit un nombre largement supérieur aux nécessités de ce type de commerce ! (Cochoy, 2014). Nous faisons ici le pari que cette histoire, *a priori* très éloignée des données numériques contemporaines, est au contraire éminemment utile pour les comprendre : il se pourrait bien que la promotion des *big data* sophistiquées d'aujourd'hui fonctionne selon la même logique que les *big cabas* très prosaïques de jadis, c'est-à-dire sur le registre des promesses de vente qui font vendre, et des arguments associés.

1.5. Big Government

Toutefois, avant d'en venir aux *big data*, il nous faut encore explorer d'autres liens de parenté qu'elles entretiennent avec des éléments plus anciens de l'expertise commerciale outillée. Notamment, le lien de l'expertise marchande avec le gouvernement, moins connu, est lui aussi crucial. L'abondante littérature actuelle sur le « New Public Management » présente comme une nouveauté certaine l'importation du savoir des affaires comme moyen de rénover l'action publique, même si cette importation a bien sûr donné lieu à des formes de réappropriation subtiles bien décrites par Philippe Bezes (2009). Pourtant, à l'origine, c'est une relation complètement inverse qui a prévalu : dans les années 1930, ce sont les innovations internes aux modes de gouvernement traditionnels qui ont servi de modèle au développement des techniques de gestion privée, et notamment du marketing. Ainsi, il se pourrait bien que ce que l'État actuel croit devoir au monde du privé ait en fait d'abord été inventé par lui-même

avant d'être recyclé par le monde des affaires ! L'éditorial qui ouvre le tout premier numéro de *The American Marketing Journal* (qui deviendra le *Journal of Marketing*, principale revue du domaine, deux ans plus tard) ne saurait mieux en témoigner, puisqu'il arrime, dès la première ligne de son tout premier texte, le sort du marketing universitaire et managérial aux innovations du gouvernement Roosevelt, à une époque où la maîtrise des marchés constitue un enjeu vital comme jamais, puisqu'il ne s'agit ni plus ni moins que de trouver les moyens de surmonter la Grande Dépression qui détruit l'Amérique tout entière :

> Alors que ce premier numéro part sous presse, la National Recovery Administration et l'Agricultural Adjustment Administration font des efforts audacieux pour hâter le retour de la prospérité. […] L'Administration pense que la stabilité de l'emploi et des salaires adéquats est de la première importance pour fournir un marché de masse à notre production de masse, et la nation est unie dans un grand effort concret visant à donner à cette conception une mise en œuvre universelle. Les résultats seront étroitement observés par les cadres du marketing, qui consacreront une attention croissante aux données et aux horaires comme indices des possibilités de vente. […] Le but de *The American Marketing Journal* est tout d'abord de présenter un matériau valable qui sera intéressant pour ceux qui ont la charge des opérations de marketing dans les entreprises privées. (*The American Marketing Journal*, 1934)

Ici et pour la première fois, politique publique et gestion du marché sont clairement confondus ; l'État et les entreprises partagent l'objectif commun de « fournir un marché de masse à notre production de masse ». L'éditorial présente le marketing non pas comme une préoccupation entrepreneuriale, gouvernée par des motifs étroitement locaux et privés, mais comme un enjeu de salut public national. Surtout, il reconnaît que l'action du gouvernement en la matière introduit une nouveauté radicale : grâce à l'interventionnisme du gouvernement Roosevelt, la conjoncture économique n'apparaît plus comme une externalité dont les autorités peuvent au mieux pallier les effets, mais comme un système d'ensemble qu'il est enfin possible de comprendre et de piloter. Comment ? Grâce à la collecte et à l'usage systématiques de données massives, bref de *big data* avant la lettre, recueillies notamment dans le cadre du premier recensement national de la distribution en 1930 (Brown, 1951, p. 61) : pour le marketing naissant, il s'agit bien d'utiliser ce type de « données » comme « indices des possibilités de vente[3] ».

En fait, le procès en paternité, c'est-à-dire la volonté de tracer l'origine publique ou privée des innovations, souvent guidé par un souci moral implicite croyant pouvoir distinguer entre bonnes innovations (au service du bien public) et mauvaises innovations (à la solde des intérêts privés), est trompeur, d'une part parce que, comme on vient de le voir, la poursuite de l'un de ces objectifs peut favoriser l'autre (et inversement), et d'autre part parce que les innovations en cause, loin d'appartenir à un domaine particulier de la vie sociale plutôt qu'à un autre, sont-elles mêmes l'expression de progrès techniques sous-jacents qui ne sont assignables à aucun d'entre eux, mais qui les traversent tous simultanément. Ce que nous apprend la parenté oubliée entre management public et marketing – dans cet ordre –, c'est que l'invention de la politique économique d'inspiration keynésienne, et donc le passage du « laissez-faire » de tradition libérale à l'interventionnisme public (ou privé) plus progressiste,

3. À son niveau, *Progressive Grocer* recourt à la même rhétorique : le magazine met en scène sa capacité à se développer à l'époque où le marché se contracte.

loin d'être réductible à une orientation idéologique particulière, est surtout la conséquence de l'invention d'une série de dispositifs techniques qui émergent, s'améliorent ou s'articulent à l'époque, tels les outils comptables, les machines mécanographiques et le calcul statistique (Desrosières, 1993 ; Didier, 2009) : sans ces moyens, aucune vision globale de l'économie n'aurait été possible, et par conséquent, aucune action de contrôle des marchés n'aurait eu de sens, tant il est vrai qu'il est vain, voire dangereux, d'agir à l'aveugle sur un monde que l'on est incapable de se représenter. Inversement, à la fin des années 1930, les nouvelles données et techniques de traitement fournissent les prothèses oculaires nécessaires à la vision (et à l'élaboration) du monde macroscopique, de sorte que tous les acteurs jadis – sinon aveugles, du moins terriblement myopes, quelles qu'aient pu être leur position et leur orientation antérieures – eurent à cœur d'exercer leurs nouvelles facultés perceptives et de se mettre, grâce à elles, à avancer et à agir, chacun au mieux de ses objectifs et de ses intérêts, comme le montrent en particulier l'exemple des compagnies d'assurance et la façon dont ces dernières ont consommé les techniques et les dispositifs de calcul disponibles à l'époque pour calculer leurs primes, développer leurs affaires et assurer leurs profits (Yates, 2006).

1.6. Big Science

Les avancées du « Big Business » et du « Big Government » dans la production et la maîtrise de données en masse étaient donc sous-tendues par l'adoption d'innovations techniques plus transversales et donc moins « assignables » à telle ou telle sphère. Ces innovations étaient elles-mêmes fondées sur une série de développements scientifiques majeurs ajoutant la « Big Science » (Galison, 1994) comme troisième volet du triptyque qui sert de clé de voûte à l'avènement du marketing scientifique, comme (avant-)dernier avatar des prétendants à la victoire au jeu de la main chaude, avant que les *big data* contemporaines ne vinssent à leur tour, comme nous le verrons, le concurrencer[4].

Les techniques d'analyse de données et de gouvernement des organisations et groupes sociaux d'après l'intelligence de ces données interviennent comme point d'articulation entre une série de contributions scientifiques, technologiques et institutionnelles qui les ont précédées ou accompagnées : citons, pêle-mêle, les premières enquêtes par questionnaire à finalité commerciale menées par le psychologue Harlow Gale à l'université du Minnesota dès la fin du XIX^e^ siècle (Eighmey et Sar, 2007), la création du Bureau of Business Research de l'université de Harvard en 1911 (Jones et Monieson, 1990), la mise au point du sondage par échantillon représentatif et autres techniques de *survey research* (Converse, 1987), le calcul probabiliste pour les compagnies d'assurance (Yates, 2006), les statistiques publiques pour le gouvernement (Desrosières, 1993 ; Didier, 2009) et, dans l'après-guerre, la reconversion civile de la recherche opérationnelle et l'essor des sciences du comportement, qui ont servi

4. Nous avions imprudemment présenté la science du marketing comme la « dernière main » dans le jeu de la main chaude (Cochoy, 1999). Nous réalisons aujourd'hui qu'il aurait été sage de préciser qu'il fallait entendre « la dernière » main comme « la plus récente » dans le cadre temporel de l'étude que nous menions à l'époque, et surtout pas comme « l'ultime » main dans un jeu qui s'est poursuivi au-delà et qui semble bien ne devoir jamais s'arrêter.

de double fondement à un puissant mouvement de quantification et de « scientifisation » du marketing, à grand renfort d'économétrie, de modélisation, d'expérimentations de laboratoire et d'analyses multivariées, etc. (Cochoy, 1999). Grâce à ces évolutions, le marketing scientifique développé dans les grandes universités et les cabinets de conseil américains pouvait prétendre asseoir la connaissance et le contrôle des marchés sur l'auscultation directe des clientèles, et court-circuiter ainsi la médiation de l'ensemble des acteurs de terrain qui les avaient précédés. Avec le marketing scientifique, on n'en est peut-être pas encore à l'ère du Big Data, mais l'on comprend que la maîtrise des marchés est déjà une grosse affaire de données.

2. Le marché des *big data* : des promesses à leur performativité

Le mouvement du Big Data que nous connaissons actuellement s'inscrit dans la continuité directe de cet effort de quantification et de scientifisation du marketing. À l'instar du marketing de jadis, les outils d'analyse des *big data* se présentent comme un ensemble d'applications marchandes qui doivent permettre aux *marketers* de terrain de développer différents outils d'aide à la décision plus ou moins automatisés censés assister leurs clients dans l'identification des leurs (les consommateurs), selon la bonne vieille logique du marché des dispositifs marchands.

Ces applications concernent quatre grands domaines d'innovations : 1) le recueil, le traitement et le stockage des données ; 2) la visualisation des données ; 3) l'automatisation de la décision ; 4) la mesure de performance. Ces quatre domaines renvoient à différents types de rhétoriques commerciales qui constituent autant de promesses quant aux avantages des *big data* pour la gestion de la relation client. Notons que ces quatre domaines d'innovation doivent être compris dans leur ensemble puisqu'ils composent *in fine* un processus de documentation marchande (Vayre, 2014). Ce processus n'a plus vraiment pour finalité de fournir un marché de masse à une production de masse comme au temps du Big Business, mais plutôt d'associer une offre de masse à une demande largement diversifiée.

Les technologies marchandes du Big Data sont donc adossées à une promesse d'intelligence technique supposée trouver le moyen de communiquer des informations pertinentes à une masse de consommateurs considérés comme très hétérogènes. Dans un contexte de *self-marketing* où les consommateurs peuvent construire leurs propres parcours informationnels en mobilisant, à leur convenance, les points de contact qu'ils ont à leur disposition (Cochoy, 2011), la grande promesse des dispositifs marchands du Big Data porte sur la possibilité de canaliser les activités de recherche et/ou de découverte d'informations sur les clients de façon à pouvoir garder la main sur le marché (Coll, 2014).

2.1. La promesse du recueil, du traitement et du stockage des données

Dans le domaine de la relation commerciale, tout projet d'outil mobilisant les *big data* commence par un problème de recueil et/ou de stockage et/ou de traitement des données. Le mouvement du Big Data que nous connaissons aujourd'hui est une forme de prolongement de celui, plus ancien, de la traçabilité, qui consistait à accumuler des masses de

données parfois considérables à des fins de contrôle qualité et/ou juridique (Torny, 1998). Pendant longtemps, en raison notamment de leur organisation « en silo », ces données ne pouvaient pas faire l'objet d'une exploitation systématique et il était impossible de passer de la traçabilité à l'intelligence des traces, ou « mappabilité » (Cochoy et Terssac, 2000). C'est précisément ce passage que promettent les grands acteurs de l'offre sur le marché des technologies *big data*, en s'appuyant sur le développement de bases de stockage qui permettent de mettre en relation des grands ensembles de données recueillies à diverses occasions et/ou à diverses fins, et de construire à partir de là une forme d'expertise que nous pourrions nommer « génie des traces ».

Ainsi, des entreprises comme Microsoft, Statistical Analysis System (SAS), Quartet FS, International Business Machine (IBM) ou encore Hewlett-Packard (HP) mettent à disposition des professionnels du marché de nouvelles architectures de stockage et de traitement dites « distribuées » et *in-memory* censées permettre de dépasser les silos de données pour procéder à leur traitement en temps (quasi) réel. À l'image de la base de données HANA développée par Systems Applications and Products (SAP), ces nouvelles architectures de stockage et de traitement sont présentées comme une véritable rupture dans le domaine de la gestion des données qui permettrait d'améliorer globalement la performance de l'entreprise[5].

Ces innovations s'appliquent au stockage et au traitement des données autant internes qu'externes à l'entreprise. Certains acteurs proposent ainsi des dispositifs ayant pour fonction d'enrichir les bases de données internes aux entreprises en allant puiser dans l'immense univers des données externes que constituent les *open data* ou encore la totalité de l'Internet. C'est par exemple le cas d'Octopeek, qui a développé une solution destinée à enrichir les bases de données clients au moyen d'éléments puisés à l'extérieur.

À travers son système d'« enrichissement à 360° » des données clients, Octopeek propose en effet une solution de « *profiling* par l'email ». Cette solution a fait l'objet de trois ans de recherche-développement et repose sur différentes techniques d'indexation (i.e. *crawling*), de structuration (i.e. *scraping*) et d'apprentissage statistique. Elle consiste à associer le courriel d'un client à un ensemble de données externes issues des réseaux sociaux, des *open data* (e.g. l'Insee), des bases de données des partenaires et des données disponibles sur l'ensemble de l'Internet. À partir de ces associations, l'objectif est de donner de la « valeur aux données clients » en déterminant différentes informations comme par exemple l'âge, la catégorie socioprofessionnelle, le lieu de résidence, les affinités, les intérêts et les styles de vie, les réseaux d'amis et/ou professionnels. Par conséquent, à travers ce type de solution, les bases de données distribuées et *in-memory* peuvent être considérées comme des améliorations fonctionnelles et technologiques des progiciels de gestion de l'entreprise que sont les Enterprise Resource Planning (ERP) et les outils de Customer Relationship Management (CRM). Comme nous l'avons vu, elles doivent en effet permettre de compléter et/ou de contextualiser les données issues de ces progiciels en les associant à d'autres données internes et externes à l'entreprise, mais aussi et surtout elles doivent permettre leur traitement en temps (quasi) réel.

5. Le slogan que SAP propose pour présenter son offre HANA en est une bonne illustration : « *Not just a database. A whole new approach to data. Run Better. SAP.* ».

Ces différentes promesses de recueil, de traitement et de stockage des données ne peuvent être correctement saisies qu'en relation avec celles que sous-tendent les trois autres domaines d'innovation. Le domaine d'innovation le plus intimement associé à celui des technologies de recueil, de traitement et de stockage des données est celui des technologies de visualisation. Car, si une entreprise peut trouver intéressant de recueillir, de stocker et de traiter en temps réel de grandes masses de données, c'est avant tout parce que d'autres entreprises promettent, à leur tour, d'opérer la mappabilité de ces mêmes données, c'est-à-dire de les rendre intelligibles et donc exploitables grâce à diverses techniques de visualisation.

2.2. La promesse de la visualisation des données

IBM, SAS, SAP, HP, Microsoft ou encore Quartet FS ne proposent pas seulement des solutions dans le domaine du stockage et du traitement des données, mais aussi dans celui de leur visualisation. L'importance accordée aux techniques de visualisation est telle qu'elle conduit certains acteurs comme QlikView ou encore Tableau Software à se spécialiser dans ce secteur d'innovation.

À travers son slogan « *Release your innate power of analysis. No limited views. Explore data naturally* », QlikView promet par exemple de faciliter l'exploration des données de manière à simplifier le travail d'enquête statistique au point d'en faire une activité quasi naturelle et sans limites. En ce sens, les promesses des acteurs de la visualisation des données massives mettent en scène une démocratisation de la pratique de l'enquête statistique, susceptible de contribuer au prolongement du mouvement de quantification du marketing. Par exemple, les tableaux de bord développés par QlikView ou encore Tableau Software ont pour finalité de simplifier les activités d'exploration des données afin de permettre à des non-statisticiens de naviguer à l'intérieur de bases de stockage complexes et parfois immenses sous la forme d'activités proches de la recherche d'informations sur Internet. Les interfaces de ces technologies sont en effet généralement élaborées de façon à être les plus intuitives possibles. Par exemple, les dégradés de couleurs peuvent être utilisés pour signifier l'intensité des corrélations statistiques d'une matrice, ou encore, quelques clics suffisent à la construction d'une carte, d'un tableau ou d'un graphique (cf. la technique du *drag and drop*). De plus, la simple sélection d'une zone ou d'une période définie permet généralement de créer un(e) autre carte/tableau/graphique précisant l'observation[6]. De cette façon, ces outils de visualisation permettent à leurs utilisateurs de faire varier très simplement les échelles d'analyse.

Encore une fois, la promesse de visualisation des données ne prend tout son sens que lorsqu'elle est mise en relation avec ses cousines : l'automatisation de la décision et la mesure de performance. L'exploration des données en soi n'a en effet pas vraiment d'intérêt pour l'entreprise si elle n'est pas directement associée à une stratégie d'exploitation (March, 1991 ; Mothe et Brion, 2008). Dans le cadre d'un projet *big data*, l'avantage escompté des

6. Pour une illustration de ces différentes fonctions : https://www.tableau.com/fr-fr/products/desktop#video.

technologies de visualisation des données est de permettre le développement, l'amélioration ou encore l'évaluation de la performance d'un automate décisionnel (i.e. d'une technique de *machine learning*). C'est ici que se trouve tout l'enjeu des applications marchandes *big data*, puisque ce serait avant tout à travers ces automates qu'il deviendrait possible d'améliorer la performance économique des actions marketing de l'entreprise.

2.3. La promesse de l'automatisation de la décision

Les projets *big data* appliqués à la relation client ont généralement pour but d'automatiser le système de communication d'une entreprise gérant une masse plus ou moins importante de consommateurs dans l'espoir de mieux cibler sa communication commerciale. Grâce à différents types d'algorithmes d'apprentissage statistique, un certain nombre d'acteurs proposent ainsi des automates décisionnels[7] capables de personnaliser les environnements numériques marchands des consommateurs de façon à améliorer la pertinence des informations marchandes. La promesse de l'automatisation de la décision relève clairement de la relation « (B to [B) to C] », puisqu'elle prétend permettre à ses clients (par exemple un e-commerçant) d'augmenter les taux de conversion entre visites et ventes effectives, ou encore de baisser les taux de désabonnement (c'est-à-dire d'augmenter les ventes auprès des consommateurs).

D'une manière très générale, un algorithme d'apprentissage statistique a pour fonction de trouver, de façon plus ou moins supervisée, la fonction f qui permet d'associer les entrées x et les sorties y d'un échantillon de données $S = [(x1, y1), \ldots, (xn, yn)]$ (Cornuéjols et Miclet, 2013), c'est-à-dire, par exemple, la règle qui permet d'établir un lien entre un profil de consommateur bien défini (donnée entrante x) et un type de produit particulier (donnée sortante y). En elles-mêmes, ces techniques d'apprentissage artificiel n'ont généralement pas grand-chose de nouveau du point de vue des disciplines sous-jacentes que sont les statistiques, l'intelligence et l'apprentissage artificiels. Par exemple, les séparateurs à vaste marge (i.e. : *support vector machine* ou SVM), aujourd'hui très largement mobilisés dans le domaine de l'automatisation de la décision, trouvent leurs origines dans les travaux de Vladimir Vapnik et de ses collègues (Vapnik et Lerner, 1963 ; Vapnik et Chervonenkis, 1964).

Comme l'expose bien le slogan de Tinyclues (« *great math, simple use, business impact* »), les mathématiciens et/ou les informaticiens sont conduits à s'allier aux commerçants afin de développer des technologies de *machine learning* qui permettent d'apprendre, en temps réel, les préférences des consommateurs de manière à pouvoir leur soumettre des informations

7. La notion d'automate décisionnel est quelquefois utilisée par les spécialistes de l'intelligence artificielle afin de désigner les systèmes d'apprentissage et de prise de décision qu'ils conçoivent. De prime abord, elle apparaît assez curieuse : une décision automatisée reste-t-elle une décision ? La réponse serait relativement simple si les automates étaient incapables d'apprendre à s'adapter à leur environnement. Ce qui n'est pas le cas : les agents artificiels peuvent apprendre des interactions qu'ils entretiennent avec leurs environnements afin d'améliorer leur performance. Ils peuvent ainsi décider de changer leurs modèles d'action en fonction de critères d'évaluation bien spécifiques et des informations qu'ils peuvent recueillir sur leur environnement.

sur des produits/services de consommation susceptibles de les intéresser. La promesse est claire et précise : on propose des mathématiques avancées pour une utilisation simple et un impact commercial réel. En d'autres termes, les mathématiques doivent servir les stratégies marketing en se substituant à l'intelligence du *marketer* traditionnel, augurant un nouvel avatar, bien après le libre-service, la publicité ou les études de marché, de la « vente sans vendeur » (Grandclément, 2008).

À la différences de Tinyclues, des acteurs comme Makazi développent des solutions qui doivent cette fois-ci permettre à leurs clients d'élaborer, de manière plus ou moins autonome, des automates décisionnels personnalisés en fonction des problématiques marketing qu'ils peuvent rencontrer. C'est par exemple le cas de Dataiku qui a mis en place une offre appelée Data Science Studio qui, en facilitant l'accès aux technologies du Big Data (et donc aux techniques d'apprentissage artificiel), doit permettre à n'importe quel commerçant de se lancer dans la science des données : les promesses du Big Data deviendraient alors une réalité pour tous.

Notons que l'enjeu de ces différentes innovations est considérable puisque, jusqu'ici, l'analyse marketing reposait sur le découplage entre temps du recueil et temps du traitement des données, sur le test d'hypothèse (*top-down*) ou l'identification de lois *via* l'économétrie (*bottom-up*), et sur la médiation technique des experts capables de conduire ces opérations. Désormais, grâce aux techniques de visualisation instantanée et aux automates décisionnels, le jeu de la main chaude reprend : on promet la disparition du marketing à l'ancienne, la relégation du temps des « études », que l'on remplace par une exploration immédiate et directe des traces par les clients eux-mêmes, moyennant bien sûr le détour par la nouvelle médiation du *marketer*-automate porté par les outils du Big Data.

2.4. La promesse de l'évaluation de la performance

Compte tenu du fait que tout apprentissage artificiel est généralement développé et/ou amélioré grâce à la mise en avant d'un critère de performance singulier, la promesse d'une automatisation de la décision réussie dépend étroitement de celle d'une évaluation rigoureuse de sa performance. En définitive, la rhétorique commerciale du Big Data consiste à dire qu'il est avant tout nécessaire de pouvoir mesurer et calculer la performance d'une action afin de pouvoir l'optimiser.

AT Internet propose ainsi une solution d'évaluation censée permettre de réconcilier les différents parcours des consommateurs en ligne et hors ligne de façon à mesurer l'impact des actions de communication qui ont été déployées durant l'ensemble de leurs expériences d'achat. L'intérêt de cette démarche dite « intégrée » serait de pouvoir rendre compte des phénomènes de composition plus ou moins réussis entre les actions de communication réalisées virtuellement et physiquement. C'est précisément en ce sens que les technologies de mesure de la performance recouvrent la promesse de constituer, pour les professionnels du marché, des systèmes de rétroaction leur permettant d'évaluer rigoureusement l'efficacité de leurs actions marketing. Par exemple, dans le cas d'une stratégie de communication multicanal impliquant plusieurs campagnes menées par plusieurs

agents humains et/ou artificiels, la solution proposée par AT Internet pourrait permettre de repérer les combinaisons gagnantes ou non afin d'optimiser les apprentissages et donc les décisions de ces deux types d'agents. Ce type d'approche est porteur d'une efficacité considérable : en bouclant une promesse sur la mesure effective de sa réalisation, on renforce considérablement son pouvoir performatif[8], dans la mesure où le spectacle d'une conjecture validée fonctionne comme garantie des réussites à venir : ce schéma, inauguré notamment par la célèbre campagne Myriam et ses promesses successives – demain, je promets d'enlever le haut ; le lendemain, je promets d'enlever le bas en même temps que j'exauce le premier vœu (Cochoy, 2011) – semble être devenu l'un des principaux ressorts de la vente des dispositifs marchands.

2.5. Une efficacité globale réelle mais à nuancer

Dès lors, qu'en est-il du succès des promesses effectuées par les offreurs de solutions *big data*[9] ? Une première manière d'examiner l'(in)félicité de ces discours est de s'intéresser au niveau de diffusion des dispositifs marchands du Big Data tels qu'ils sont relayés (et/ou narrés) par les acteurs du secteur eux-mêmes. Le théâtre des preuves mises en avant par une industrie pour promouvoir ses succès à venir est peut-être discutable en soi, mais aussi consubstantiel au fonctionnement du marché des dispositifs marchands : nous avions vu, par exemple, comment la revue *Progressive Grocer* avait construit sa fortune auprès des annonceurs grâce au tour de passe-passe consistant à exhiber la progression de son chiffre d'affaires concernant la vente de son espace publicitaire comme la preuve de l'efficacité commerciale de ces mêmes annonces. Le marché des *big data* reprend aujourd'hui la même rhétorique, qui consiste à exhiber des chiffres favorables comme indices de succès à venir, en vertu du schéma assez général d'après lequel la performativité d'une nouvelle promesse se renforce lorsqu'on l'accompagne de l'exhibition de l'effectivité de promesses passées (cf. *supra*). On aurait tort, ici, d'opposer rhétorique et réalité : comme toujours en matière de statistiques (Desrosières, 1993), les chiffres

8. La pragmatique linguistique a introduit la distinction entre énoncés « constatifs » et énoncés « performatifs » : tandis que l'on peut établir le caractère de vérité ou de fausseté des premiers en se référant au réel vers lequel ils pointent (« le chat est sur le paillasson »), les seconds ne sont ni vrais ni faux, mais font ou peuvent faire advenir le monde auquel ils se réfèrent (« je promets que je viendrai demain » ; « je vous déclare mari et femme » ; etc.). Cette notion a récemment fondé, sous l'impulsion de Michel Callon (1998), puis d'autres auteurs, un important renouvellement de la sociologie économique, dans la mesure où elle permet d'écarter la vaine querelle qui prédominait jusqu'alors sur la vérité ou la fausseté des théories économiques et des sciences de gestion (leur caractère constatif) pour s'intéresser à la façon dont leurs énoncés contribuent, ou non, sous certaines conditions, à transformer le monde dont elles traitent.
9. Nous adoptons ici un angle de vue bien spécifique qui consiste à rechercher comment les discours des offreurs de solutions *big data* participent à orienter la pratique des marchands. Ce point de vue est bien entendu très discutable. Par exemple, il ne permet pas de rendre correctement compte du rôle des éléments matériels dans la dynamique du marché des dispositifs marchands *big data*. Or, il est clair que cette dynamique est façonnée par de nombreux enjeux sociotechniques. Et, les acteurs de l'offre élaborent leurs promesses en fonction de ces enjeux qu'ils connaissent souvent assez bien. Nous reviendrons toutefois sur ce point afin de souligner comment ce jeu promesses/pratiques peut former une sorte de boucle itérative qui comporte un certain nombre de risques sur le plan socio-économique.

mobilisés sont autant réels (ils ne sont pas inventés, mais renvoient à des référents et à des mesures objectifs) que construits (ils sont définis, mis en scène et sélectionnés selon l'intérêt bien compris de ceux qui les produisent).

Ainsi, le *Guide du Big Data 2013/2014* souligne que ce marché semble susciter un réel intérêt du côté des fournisseurs de prestations puisqu'il connaît un taux de croissance annuel de 31,7 % selon l'International Data Corporation (IDC). Toutefois, la situation est bien plus contrastée lorsque l'on se place du côté des clients. Par exemple, le *Big Data Index EMC/IDC* (2012) estime que seulement 10 % des entreprises françaises mobilisent les technologies du Big Data alors que plus de 70 % d'entre elles n'ont encore jamais projeté d'adopter, ne serait-ce qu'à titre de thème de réflexion, ces mêmes technologies[10].

Cette performativité relative des promesses associées à la commercialisation des applications marchandes *big data* peut être expliquée de plusieurs façons. Les avancées contemporaines de la linguistique pragmatique (Austin, 1962 ; Searle, 1969 ; Grice, 1975 ; Sperber et Wilson, 1986) montrent que, du point de vue de l'interlocuteur, la félicité d'un énoncé dépend des contextes objectif et subjectif de son énonciation. Or, il apparaît que les promesses des offreurs ne sont pas toujours perçues comme pertinentes par un certain nombre de demandeurs. À en croire la presse spécialisée, les entreprises n'ont pas nécessairement les ressources économiques, techniques, organisationnelles ou encore culturelles qui pourraient leur permettre d'envisager un projet de développement *big data*. Autrement dit, il apparaît encore aujourd'hui que, pour une bonne partie des entreprises, les technologies marchandes du Big Data recouvrent un coût d'intégration économique et/ou sociotechnique potentiellement trop élevé en rapport à leur utilité. Les discours ne suffisent pas à soutenir les pratiques, pour la bonne raison que la mise en œuvre des concepts requiert ici de lourds investissements matériels dont le coût vient modérer les promesses marchandes.

Néanmoins, le rapport d'enquête publié en 2013 par Tata Consultancy Services (TCS) propose un angle de vue un peu différent[11], qui pourrait suggérer que la performativité globale des promesses des offreurs de solutions marchandes *big data* n'est peut-être pas si faible qu'elle n'y paraît. En effet, sur les 1217 entreprises qui ont fait l'objet d'une enquête par TCS durant les mois de décembre 2012 et janvier 2013, un peu plus de la moitié (50,75 %) ont lancé une ou plusieurs initiatives dans le domaine du Big Data (68 % des firmes américaines, 51 % des firmes latino-américaines, 48 % des firmes européennes et 39 % des firmes d'Asie-Pacifique). Aussi est-il important de noter que les compagnies sur lesquelles TCS a mené des enquêtes sont des entreprises qui bénéficient des services informatiques proposés par TCS. En ce sens, ces différentes entreprises sont *a priori* prédisposées (sur les plans économique, technique, organisationnel et culturel) à mettre en place un projet d'intégration d'innovations issues du domaine

10. La première version du *Big Data Index EMC/IDC* a été réalisée en 2012 et a pour objectif de mieux évaluer la perception que les entreprises françaises ont du Big Data et surtout de mieux mesurer les réalités des initiatives prises dans ce domaine. Le rapport repose sur une enquête conduite en France de juin à juillet 2012 auprès de 160 entreprises de plus de 200 salariés.
11. Cf. *The Emerging Big Returns on Big Data. A TCS 2013 Global Trend Study*.

de l'informatique organisationnelle et/ou commerciale. Par conséquent, il apparaît que les promesses des offreurs sur le marché des dispositifs marchands *big data* sont, dans ce cas de figure, perçues comme pertinentes.

Dans l'ensemble, pour être faible et fragile, la performativité des promesses des offreurs sur le marché des dispositifs marchands *big data* est réelle. Par exemple, sur les 643 entreprises ayant fait l'objet d'enquêtes par TCS qui ont réalisé un projet de développement *big data*, 81,5 % estiment avoir effectivement amélioré le procès décisionnel de l'entreprise[12] (soit 77 % des entreprises américaines, 80 % des entreprises européennes, 83 % des entreprises d'Asie-Pacifique et 86 % des entreprises latino-américaines). En outre, 76 % des entreprises sur lesquelles TCS avait enquêté et ayant adopté des technologies marchandes du Big Data affirment avoir bénéficié, en 2012, d'un retour sur investissement positif.

2.6. Une forte performativité locale

À un niveau plus localisé, la performativité des promesses des offreurs de dispositifs marchands *big data* est plutôt importante. Pour commencer, l'enquête de TCS (2013) montre que le secteur du travail relationnel marchand capte la majeure partie des investissements réalisés dans le cadre d'un projet de développement *big data* (43,5 %). Plus précisément, 13,3 % sont effectués dans le domaine du service à la consommation, 15 % dans celui du marketing et 15,2 % dans celui de la vente. Aussi, lorsqu'on regarde les pourcentages moyens attendus des retours sur investissement dans le domaine des technologies *big data*, on s'aperçoit que, bien que les services à la consommation (55,9 %) et la vente (54,3 %) soient en dessous de la logistique/distribution (78,1 %) et de la finance (69 %), ces derniers restent légèrement au-dessus de la moyenne générale (54 %) et font partie des plus élevés. Notons néanmoins que dans le cas du marketing, les espérances sont moins élevées puisque le pourcentage moyen attendu des retours sur investissement dans le domaine des technologies *big data* est de 41,4 % (i.e. la dernière position). Au total, ces résultats restent toutefois ambivalents. D'un côté, ils traduisent une mise en œuvre certaine, même partielle, de ces technologies ; de l'autre, leur mise en scène par l'entreprise TCS reprend la rhétorique de la presse professionnelle, qui consiste à organiser le théâtre du succès des thématiques dont elle est porteuse pour encourager le trafic ultérieur et développer ses propres affaires. Ici, la question n'est pas tellement de mettre en doute les chiffres présentés dans l'enquête de TCS (qui sont pour la plupart recueillis à travers des questions factuelles), mais plutôt le choix des thématiques abordées et celui des résultats publiés. En ce sens, l'enquête de TCS illustre la performativité relative des promesses réalisées par les offreurs de solutions *big data* qui n'apparaît pas vraiment contestable telle qu'elle. Par contre, le rapport de TCS n'aborde pas un pan gigantesque

12. TCS ne précise pas directement la question posée aux enquêtés. Néanmoins, le rapport stipule qu'il s'agissait ici de demander aux participants si leur(s) initiative(s) de développement *big data* leur avai(en)t permis d'améliorer leur prise de décision dans le domaine des affaires.

de questions qui pourraient permettre de compléter et/ou de nuancer ces résultats. Par exemple, comme nous l'avons déjà signalé, dans le cas des systèmes de personnalisation des environnements numériques marchands (cf. l'automatisation de la décision), les acteurs mettent généralement en avant les taux de conversion ou de désabonnement. Or, il existe de nombreux indicateurs de mesure de désorientation qui ne sont pas pris en compte et qui pourraient permettre de discuter l'efficacité et la pertinence de ce type de solution du point de vue des usages qu'en font les consommateurs finaux : comme toujours, il y a loin entre la vente des dispositifs marchands et l'efficacité de leur fonctionnement.

En outre, en examinant plus en détail le rapport d'enquête de TCS (2013), on s'aperçoit que les quatre types de promesses que nous avons présentés en amont participent pleinement à orienter les attentes et les projets des entreprises. En effet, dans le secteur du travail relationnel marchand et à travers une liste constituée de plus d'une soixantaine de *challenges* et de thématiques pouvant constituer de potentielles sources de bénéfices, les entreprises enquêtées par TCS disent s'intéresser, de façon générale, à ceux-ci (tableau 1).

Ainsi, les offreurs semblent participer à la « performation » des projets et des problématiques que recouvre l'implémentation, au sein de l'entreprise, des dispositifs marchands *big data*. Cette façon de faire est porteuse d'un risque non négligeable résidant dans l'instauration d'une sorte de processus de bouclage limitatif entre l'offre et la demande fondé sur l'enfermement dans deux séquences itératives : d'une part, l'offre fait des promesses qui orientent les usages de la demande ; d'autre part, les usages de la demande sont examinés par l'offre afin d'élaborer de (nouvelles) promesses. En instituant des formes d'ajustement incrémentales sur le marché des dispositifs marchands *big data*, il se pourrait bien que ce processus contribue à favoriser le manque d'imagination qui semble caractériser, comme l'a souligné Christophe Benavent lors du salon « Big Data » en 2014, les quatre domaines d'innovation présentés en amont. À tout le moins, il apparaît que le processus de documentation que recouvrent les technologies marchandes *big data* renvoie, du point de vue des sciences humaines et sociales, à un ensemble de problématiques théoriques et méthodologiques qui sont plus ou moins ignorées par leurs concepteurs et leurs utilisateurs. Pourtant, la considération de ces problématiques pourrait avoir trois conséquences interdépendantes : permettre à l'ensemble des acteurs du marché des dispositifs marchands *big data* de se distancier des enjeux économiques ; améliorer les qualités scientifique et éthique des dispositifs marchands *big data* ; favoriser la diversité des solutions et des usages des dispositifs marchands *big data*.

Promesse de recueil, de stockage et de traitement des données	• Amélioration de la rapidité du traitement d'une masse de données toujours plus nombreuses et variées ; • Possibilités de dépassement des silos organisationnels afin de pouvoir partager efficacement les données.
Promesse de visualisation des données	• Amélioration et présentation des analyses issues des *big data* pour aider la prise de décision ; • Identification de la valeur des consommateurs ou encore du risque d'attrition ; • Amélioration de la finesse de la granularité des techniques de segmentation ; • Définition des offres promotionnelles, des stratégies marketing et des messages les plus performants ; • Identification des centres d'intérêts des visiteurs et/ou des pages les plus/moins visitées en fonction de leurs parcours de navigation ; • Identification des modèles (i.e. patterns) de plaintes des clients.
Promesse d'automatisation de la décision	• Prédiction des comportements des consommateurs ; • Articulation des offres promotionnelles, stratégies marketing et formes des messages les plus performantes ; • Personnalisation des résultats de recherche sur les sites marchands ; • Identification des modèles (i.e. *patterns*) de plaintes des clients.
Promesse de l'évaluation de la performance	• Amélioration de l'efficacité des campagnes marketing et des canaux de communication ; • Définition et articulation des offres promotionnelles, stratégies marketing et formes des messages les plus performantes.

Tableau 1. Récapitulatif des centres d'intérêts des entreprises enquêtées par TCS.

Conclusion

Les *big data* instaurent un grand marché du traitement des données marchandes massives. Notre exposé peut se lire à rebours : il s'achève par la mise en évidence d'une innovation importante, assise sur des technologies numériques, et sur des promesses réelles, mesurables à l'aune de leur efficacité et de l'extension du marché des outils sous-jacents. Mais au vu de notre récit inaugural, on comprend que cette innovation s'appuie sur une série d'éléments assez anciens, comme le traitement des données de consommation du côté du marketing, et les méthodes statistiques de classification du côté du *machine-learning*. Le rapprochement de l'actualité et de l'histoire permet de pointer, dans l'entre-deux, une série d'éléments novateurs comme l'exploitation secondaire des traces numériques, l'émergence d'une analyse en temps réel du marché, la promotion d'une pratique directe et profane de l'étude, c'est-à-dire d'un marketing (presque) sans *marketers*, qui complète

la vente sans vendeurs du libre-service grâce à la substitution (au moins rhétorique) du génie logiciel aux *marketers* de jadis. Il y a là un réel enjeu social, qui permet en quelque sorte, pour reprendre une métaphore que nous avons longtemps filée, à la nouvelle main de silicium de prendre l'avantage sur la « main » de chair ou de papier du marketing dans l'ambition multiséculaire de maîtrise des marchés. Reste la question du dimensionnement des solutions proposées aux pratiques réelles, de l'adéquation des outils à leur demande. Quelle demande ? La catégorie du « (B to [B) to C] » est ici très utile, dans la mesure où elle nous rappelle combien la dynamique du marché des dispositifs marchands repose sur la promesse faite aux professionnels quant au pouvoir qu'auraient ces dispositifs d'améliorer ou de multiplier les ventes auprès des consommateurs finaux. On retrouve alors, derrière les atours modernistes du Big Data, le théâtre habituel de l'efficacité supposée des technologies de manipulation… qui ne manipulent de façon certaine que les candidats au métier de manipulateur.

Note des auteurs : L'écriture de ce chapitre a bénéficié du soutien du Swedish Research Council (project Digcon: Digitalizing consumer culture, Grant number 2012-5736).

Références

Austin J.-L. (1962), *How to do Things with Words*, Oxford, Oxford University Press.

Bezes P. (2009), *Réinventer l'État. Les réformes de l'administration française (1962-2008)*, Paris, PUF.

Bruno I. (2012), « Quand s'associer, c'est concourir . Les paradoxes de la "coopétition" », *in* Cochoy F. (dir.), *Du lien marchand. Essai(s) de sociologie économique relationniste*, Toulouse, Presses universitaires du Mirail, p. 54-78.

Brown G.H. (1951), « What economists should know about marketing », *Journal of Marketing*, vol. 16, nº 1, p. 60-66.

Callon M. (1998), *The Laws of the Markets*, Oxford, Blackwell.

Chandler Alfred D. Jr. (1988), *La Main visible des managers : une analyse historique*, Paris, Economica [1977].

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

Cochoy F. (2011), *Sociologie d'un « curiositif » : smartphone, code-barres 2D et self-marketing*, Lormont, Éditions Le Bord de l'eau, coll. « Mondes marchands ».

Cochoy F. (2014), *Aux origines du libre- service. Progressive Grocer (1922-1959)*, Lormont, Éditions Le Bord de l'eau, coll. « Mondes marchands ».

Cochoy F. (2015a), « Afterword. Concerned markets: facing the future, beyond "interested" and "contested" markets » *in* Harrison D., Geiger S., Kjellberg H. et Mallard A. (éd.), *Concerned Markets. Economic Ordering for Multiple Values* , Cheltenham, Edward Elgar Publishing.

Cochoy F. et Terssac (de) G. (2000), « Au-delà de la traçabilité : la mappabilité. Deux notions connexes mais distinctes pour penser les normes de management », *in* Serverin E. et Berthoud A. (dir.), *La Production des normes entre État et société civile*, Paris, L'Harmattan, p. 239-249.

Coll S. (2014), *Surveiller et récompenser : les cartes de fidélité qui nous gouvernent*, Zurich/Genève, Seismo, coll. « Terrains des sciences sociales ».

Converse J. (1987), *Survey Research, United States Roots and Emergence, 1890-1960*, Berkeley, CA, University of California Press.

Cornuéjols A. et Miclet L. (2013), *Apprentissage artificiel. Concepts et algorithmes*, Paris, Eyrolles.

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

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

Galison P.L. (1994), *Big Science: The Growth of Large Scale Research*, Stanford, Stanford University Press.

Grandclément C. (2006), « Wheeling food products around the store… and away: the invention of the shopping cart, 1936-1953 », papier présenté au *Food Chains Conference: Provisioning, Technology, and Science*, Hagley Museum and Library, Wilmington, Delaware, 2-4 novembre : http://www.csi.ensmp.fr/Items/WorkingPapers/Download/DLWP.php?wp=WP_CSI_006.pdf.

Grandclément C. (2008), *Vendre sans vendeurs : sociologie des dispositifs d'achalandage en supermarché*, thèse pour obtenir le grade de docteur de l'École de mines de Paris, spécialité « Socio-économie de l'innovation ».

Grice P.H. (1975), « Logic and conversation », *in* Cole P. et Morgan J.-L. (éd.), *Syntax and Semantics 3: Speech Acts*, New York, Academic Press, p. 41-58.

Eighmey J. et Sar S. (2007), « Harlow Gale and the origins of the psychology of advertising », *Journal of Advertising*, vol. 36, n° 4, p. 147-158.

Hirschman A.O. (1980), *Les passions et les intérêts, justifications politiques du capitalisme avant son apogée*, Paris, Presses universitaires de France.

Jones D.G.B. et Monieson D. D. (1990), « Early development of the philosophy of marketing thought », *Journal of Marketing*, vol. 54, n° 1, p. 102-113.

March J.G. (1991), « Exploration and exploitation in organizational learning », *Organization Science*, vol. 2, n° 1, p. 71-87.

Martin D. (2005), *Les Options fondamentales de la finance : domestication sociologique d'un produit financier, thèse pour le doctorat de sociologie*, Toulouse, université Toulouse 2.

Mothe C. et Brion S. (2008), « Innovation : exploiter ou explorer ? », *Revue française de gestion*, n° 187, p. 101-108.

Polanyi K. (1983), *La Grande Transformation, aux origines politiques et économiques de notre temps* [1944], Paris, Gallimard, Bibliothèque des sciences humaines.

Rochet J.-C. et Tirole J. (2003), « Platform competition in two-sided markets », *Journal of the European Economic Association*, vol. 1, n° 4, p. 990-1029.

Searle J.R. (1969), *Speech Acts*, Cambridge, Cambridge University Press.

Sperber D. et Wilson D. (1986), *Relevance. Communication and Cognition*, Oxford, Blackwell.

Strasser S. (1989), *Satisfaction Guaranteed. The Making of the American Mass Market*. New York, Pantheon Books.

Tedlow R.S. (1990), *New and Improved. The Story of Mass Marketing in America*, New York, Basic Books.

Torny D. (1998), « La traçabilité comme technique de gouvernement des hommes et des choses », *Politix*, n° 44, 4e trimestre, p. 51-75.

Yates J. (2006), « How business enterprises use technology: extending the demand-side turn », *Enterprise and Society*, vol. 7, n° 3, p. 422-455.

Vapnik V. et Lerner A. (1963), « Pattern recognition using generalized portrait method », *Automation and Remote Control*, vol. 24, p. 774-780.

Vapnik V. et Chervonenkis A. (1964), « A note on one class of perceptrons », *Automation and Remote Control*, vol. 25.

Vayre J.-S. (2014), « Manipuler les données. Documenter le marché. Les implications organisationnelles du mouvement Big Data », *Les Cahiers du numérique*, vol. 10, n° 1, p. 95-125.

Gouverner, échanger, sécuriser

Les *big data* et la production du savoir numérique

Bernard E. Harcourt
Professeur de droit et de sciences politiques à l'université de Columbia et directeur d'études à l'EHESS

Commençons avec trois anecdotes relatives aux arts de gouverner, d'échanger et de sécuriser[1]. La première a eu lieu en 2009, et impliqua une importante base de données contenant un demi-million de numéros de téléphones portables à Cuba. L'affaire impliqua également quelques officiels et sous-traitants de l'Agence américaine pour le développement international (US Agency for International Development/USAID), une agence gouvernementale des États-Unis qui distribue des milliards de dollars d'aide et d'assistance humanitaire à des pays dans le besoin[2]. Les numéros de téléphones portables avaient été subrepticement donnés par un employé de Cubacel, l'opérateur étatique de téléphonie cubain, à un ingénieur cubain qui vivait alors en Espagne ; cet ingénieur cubain transmit par la suite cette base de données – et ce « à titre gratuit », si l'on en croit les documents examinés par l'Associated Press (AP) – aux officiels de l'USAID ainsi qu'à une société commerciale de Washington D.C., la Creative Associates International[3] (laquelle avait gagné des millions de dollars en contractant avec l'USAID).

Une manager de Creative Associates se montra alors, dira-t-on, « créative ». Avec son frère, qui résidait au Nicaragua, elle eut l'idée d'envoyer des SMS en masse depuis plusieurs pays aux utilisateurs de téléphones portables cubains afin de contourner le contrôle strict qu'opère l'État cubain sur Internet et de commencer une forme de réseau social sur le modèle de Twitter. L'idée était d'essayer discrètement de mettre en place, *ex nihilo*, une plate-forme entière pour les Cubains, un « Twitter cubain », avec à terme l'objectif de susciter une opposition politique. Comme le rapporte l'Associated Press (AP), après avoir examiné en détail plus de 1 000 pages de documents :

> Les documents montrent que le gouvernement des États-Unis avait projeté de construire une base d'abonnés par le biais de « contenus non-controversés » : des messages d'informations sur le football, de la musique, et des mises à jour sur les ouragans. Ensuite, lorsque le réseau aurait atteint une masse importante d'abonnés, peut-être des centaines de milliers, les opérateurs introduiraient alors du contenu politique destiné à inciter les Cubains à organiser des « foules intelligentes » – des rassemblements de masse qui pourraient d'une seconde à l'autre déclencher un Printemps cubain ou, comme le dit un document de l'USAID, « renégocier l'équilibre du pouvoir entre l'État et la société[4] ».

1. Je suis très reconnaissant à Pierre-Michel Menger d'avoir organisé ce travail collectif et la belle journée d'étude en juin 2014 ; à Daniel Henry pour son assistance exceptionnelle en recherche ; ainsi qu'à Julien Larregue et à Sacha Raoult pour leur traduction.
2. Butler D., Gillum J. et Arce A., « US secretly created "Cuban Twitter" to stir unrest », *Associated Press*, 4 avril 2014 : http://bigstory.ap.org/article/us-secretly-created-cuban-twitter-stir-unrest.
3. *Ibid.*
4. *Ibid.*

Ils donnèrent à ce réseau le nom de « ZunZuneo », ce qui signifie, en argot cubain, « tweet de colibri ». En mars 2011, il y avait environ 40 000 abonnés à ce système, mais aucun d'eux n'avait alors pris conscience que ce réseau social avait été créé, soutenu et nourri par les employés de l'USAID. Aucun d'eux ne s'était rendu compte que la participation de chaque membre du réseau était profilée par l'USAID, et ce afin de déterminer les tendances politiques de chacun. Aucun d'eux n'a soupçonné que cette messagerie avait pour but de les politiser. Non seulement l'USAID a fondé « un système labyrinthique de sociétés-écrans en utilisant un compte bancaire ouvert aux îles Caïmans, et a recruté des exécutants qui ignoraient tout des liens entre la société et le gouvernement des États-Unis », selon l'enquête de l'AP, mais elle a également fait mettre en place par le biais d'une entreprise britannique une société en Espagne afin de diriger ZunZuneo. L'USAID a également mis en place un site internet partenaire du service d'envoi de SMS, afin que les utilisateurs de téléphones portables puissent souscrire, émettre des réactions et envoyer des messages eux-mêmes gratuitement. Les documents révèlent la discussion qui a eu lieu sur la façon de faire paraître tout cela légitime : « Des imitations de bannières publicitaires donneront au site l'apparence d'une entreprise commerciale », peut-on ainsi y lire[5].

Une des entrepreneuses de l'USAID, Paula Cambronero, était chargé de répartir les abonnés cubains entre les catégories « pro-révolution », « apolitique » ou « anti-révolution », à partir de leurs réponses aux messages[6]. La première préoccupation du projet, selon l'AP, était de « déplacer davantage de personnes vers le camp des activistes démocratiques sans être repéré ». « Les documents de l'USAID disent que leur objectif stratégique à Cuba était de "les faire sortir de l'impasse grâce à des initiatives tactiques et temporaires, et de relancer le processus de transition vers le changement démocratique[7]" ».

ZunZuneo ferma au milieu de l'année 2012, lorsque l'argent de l'USAID s'assécha. Le sénateur Patrick Leahy du Vermont, président du Senate Appropriations Committee's State Department and Foreign Operations Subcommittee, a tenu en 2014 des audiences sur ce programme de l'USAID.

La seconde anecdote concerne l'art d'échanger. En 2004, Google lança Gmail, un service gratuit de messagerie électronique. Ce qui a rendu l'idée de Gmail si attractive pour les consommateurs, c'est que Google a fourni à ses utilisateurs une grande quantité d'espace de stockage gratuit : 1 gigaoctet pour chaque utilisateur[8]. Avant même que le produit soit mis en ligne, un grand nombre d'utilisateurs essayaient d'avoir un accès prioritaire à ce service, et certains ont même payé pour cela. Mais en échange de ce libre accès au *cloud*, les utilisateurs devaient céder à Google un peu de leur liberté : un accès-libre aux e-mails et aux pièces jointes de tous les utilisateurs – ainsi qu'à leur contenu – de même qu'un

5. *Ibid.*
6. Butler D. et Arce A., « US contractors profiled "Cuban Twitter" responses », *Associated Press*, 30 avril 2014 : http://abcnews.go.com/International/wireStory/usaid-contractors-profiled-cuban-twitter-users-23532155.
7. *Ibid.*
8. Levine Y., « The psychological dark side of Gmail: Google is using its popular Gmail service to build profiles on the hundreds of millions of people who use it », *AlterNet.org*, 31 December 2013 (première publication dans *Pando-Daily*) : http://www.alternet.org/media/google-using-gmail-build-psychological-profiles-hundreds-millions-people.

accès libre aux e-mails entrant provenant de non-souscripteurs, c'est-à-dire les e-mails de n'importe quel usager communiquant avec un utilisateur de Gmail.

Alors, avec cette plate-forme Gmail, Google scannerait automatiquement tous les mails et leur contenu afin d'adresser des publicités plus ciblées aux utilisateurs. C'est quelque chose que la plupart d'entre nous savons et que nous en sommes venus à accepter. Nous avons appris à vivre avec, c'est en quelque sorte le prix de la gratuité. Mais un journaliste d'investigation, Yasha Levine, a révélé l'ampleur de cette surveillance, bien au-delà de la publicité :

> Google ne faisait pas que scanner nos e-mails pour les mots-clés publicitaires. Il a également développé une technologie sous-jacente permettant de constituer des dossiers sophistiqués sur n'importe quelle personne qui passerait par son système de messagerie. Toute communication était sujette à une analyse linguistique poussée ; les utilisateurs étaient rattachés à leurs identités réelles par le biais des informations contenues dans leur carnet d'adresses ; les pièces jointes étaient récupérées pour les services de renseignement – cette information était ensuite recoupée avec les mails précédents et combinée avec des choses provenant d'autres services proposés par Google, de même qu'avec des sources tierces[9]…

Grâce à une fine analyse des deux brevets que Google a déposés avant de lancer le service Gmail, Levine rapporte que la compagnie utilise une large gamme de technologies afin de construire le profil de ses utilisateurs – afin de produire un savoir numérique. Ce savoir repose sur plusieurs opérations : examiner les concepts et les sujets discutés par les utilisateurs dans leurs e-mails ; analyser le contenu des sites internet que les différents utilisateurs ont visités ; collecter les informations sociodémographiques relatives aux utilisateurs, et notamment leurs revenus, leur sexe, leur ethnicité, et leur statut matrimonial ; relier cela à leurs informations géographiques ; en déduire leur profil psychologique et « psychographique », tel que leur type de personnalité, leurs valeurs, leurs centres d'intérêt et leurs attitudes ; disséquer l'historique des recherches internet de l'utilisateur ; collecter des informations à propos de n'importe quel document que l'utilisateur a vu et édité ; et étudier leurs précédents achats[10].

Le but, naturellement, est de produire des profils pour chaque utilisateur afin de cibler la publicité, d'aider à faciliter la consommation et ainsi de récolter un profit substantiel pour Google. Ce type de surveillance commerciale est devenu si important, en fait, qu'il éclipse aujourd'hui – et, comme nous allons le voir, en vient à nourrir – celui qui est conduit par l'Agence nationale de la sécurité (National Security Agency/NSA), le Quartier général des communications du gouvernement (Government Communications Headquarters/GCHQ), la Direction générale de la sécurité extérieure (DGSE), etc. Comme le suggère Levine, Gmail et d'autres services représentent « une opération massive de surveillance qui intercepte et analyse des téraoctets de trafic internet chaque jour, et les utilise alors pour construire et mettre à jour des profils psychologiques complexes de centaines de millions de personnes à travers le monde – tout cela en temps réel[11] ».

9. *Ibid.*
10. *Ibid.*
11. *Ibid.*

La troisième et dernière anecdote concerne l'art de sécuriser. L'initiative des services de renseignement britanniques avait pour nom de code *Optic Nerve*[12] (« Nerf optique »). Nous ne savons pas aujourd'hui si le GCHQ, l'agence britannique de renseignement, la conduit encore, bien qu'il y ait des preuves que c'était toujours le cas en 2012. Ce que l'on sait c'est que, pendant une période de six mois au cours de l'année 2008, l'agence britannique de renseignement a intercepté des captures d'écran des communications vidéo par webcam d'environ 1,8 millions d'utilisateurs d'Internet se servant de plates-formes de chat vidéo comme celles proposées par Yahoo Messenger[13].

À l'inverse d'autres programmes qui ne capturent que des métadonnées, Optic Nerve a pu avoir accès au contenu des communications vidéo – aux véritables images vidéo apparaissant lors de la conversation. Apparemment, le programme « téléchargeait automatiquement le contenu des communications vidéo – prenant une capture d'écran du flux vidéo toutes les cinq minutes[14] ». Selon un rapport secret dévoilé par Edward Snowden, les services de renseignement britanniques aspiraient à capturer plus d'images à un rythme plus rapide et espéraient obtenir la vidéo complète à un certain point, au moins pour les cibles des surveillances, avec l'intention d'« identifier les cibles en utilisant un logiciel de reconnaissance faciale automatique[15] ». Le GCHQ était aidé, dans ces efforts, par la NSA :

> L'information provenant des webcams était introduite dans l'outil de recherche de la NSA XKeyscore, et les recherches de la NSA ont été utilisées afin de construire l'outil qui a permis d'identifier le trafic webcam de Yahoo[16].

Apparemment, l'opération a permis de découvrir une mine d'images classées X. Un document des renseignements anglais à propos du programme déclare qu'« il semblerait qu'un nombre surprenant de personnes utilisent les conversations webcam pour montrer des parties intimes de leur corps à d'autres personnes » ; selon une analyse informelle, à peu près 7 % des images enregistrées contenaient de la « nudité non désirée » – enfin, « non désirée » par les hauts fonctionnaires des renseignements anglais[17] (encore que…). Comme le note l'Associated Press, « la collection de photographies de personnes nues soulève aussi des questions à propos des possibilités de chantage[18] ».

D'autres documents divulgués par Snowden révèlent que la NSA a étudié la possibilité d'utiliser le système de communication vidéo des consoles de jeux comme moyen d'intercepter des données : « la NSA explorait les capacités vidéo des consoles de jeu à des fins de surveillance »,

---

12. Ackerman S. et Ball J., « Optic nerve: millions of Yahoo webcam images intercepted by GCHQ », *The Guardian*, 27 février 2014 : http://www.theguardian.com/world/2014/feb/27/gchq-nsa-webcam-images-internet-yahoo ; Associated Press, « British spies intercept webcam pictures, report says », *The New York Times*, 27 février 2014 : http://www.nytimes.com/aponline/2014/02/27/world/europe/ap-nsa-surveillance-naked-pictures.html?partner=rss&emc=rss
13. Ackerman et Ball, art. cit.
14. Associated Press, « British spies intercept... », art. cit.
15. *Ibid.* ; cf. aussi Ackerman et Ball, art. cit.
16. *Ibid.*
17. Associated Press, « British spies intercept... », art. cit. ; Ackerman et Ball, art. cit.
18. Associated Press, « British spies intercept... », art. cit.

selon le *Guardian*. « Microsoft, le constructeur de la Xbox, a dû faire face à une réaction violente sur le terrain de la vie privée l'année dernière quand il est apparu que la caméra fournie avec sa nouvelle console, la Xbox One, serait toujours activée par défaut[19] ».

Chacune de ces initiatives dans les arts de gouverner, d'échanger et de sécuriser génère une mine de données additionnelles à propos des individus qui peuvent être capturées, enregistrées, reliées les unes avec les autres, connectées à d'autres, extraites, étudiées et analysées. Chacune d'entre elles peut produire une mine de données utilisables par la recherche en sciences sociales. Mais de la même manière, chacune d'entre elles peut produire une mine de données pour identifier les individus, les inciter à prendre position politiquement, exercer de petites manipulations, encourager leur consommation, stimuler ce qu'ils révèlent, les observer, les surveiller, les détecter, les prédire et les punir. Les données permettent aux gouvernements, aux multinationales, aux entreprises du quotidien, aux employeurs, aux vendeurs, aux publicitaires, à la police et aux agents d'insertion et de probation de suivre les mouvements physiques des individus, leur navigation sur Internet, de savoir ce qu'ils lisent, ce qu'ils aiment, les vêtements qu'ils portent, avec qui ils communiquent, où et comment ils dépensent leur argent…

Un nouveau savoir numérique émerge, lequel a commencé à remettre en question les lignes traditionnelles entre gouverner, échanger et sécuriser – ce qui revient à réinterroger les limites conventionnelles entre le politique, l'économique et le maintien de l'ordre. Un savoir numérique qui rend ambiguës les limites entre le commerce et la surveillance, le gouvernement et l'échange, la démocratie et l'État policier. Ce nouveau savoir numérique produit et reproduit des sujets-consommateurs qui, sciemment ou non, se laissent observer, traquer, relier et prédire au sein d'un amalgame confus de projets commerciaux et gouvernementaux. En reliant, dans les deux sens, les données sur les consommateurs aux informations gouvernementales et aux médias sociaux, ces nouvelles toiles d'informations deviennent disponibles à quiconque peut les acheter.

Comment se fait-il que les intérêts gouvernementaux, commerciaux et de sécurité aient convergé, coïncidé, mais aussi divergé sur certains points, dans cette production des *big data* ? Quels secteurs ont stimulé la production et la collecte de cette masse d'informations ? Comment ces divers projets se sont-ils alignés ou contredits ? Comment se fait-il, par exemple, que toute nouvelle technologie numérique semble faciliter la sécurité ?

Dans la suite de ce chapitre, je vais explorer ces interrogations à partir de deux dimensions. Tout d'abord, je vais exposer à grands coups de pinceaux l'évolution historique et la croissance de la production de l'univers numérique. Je vais proposer certaines catégories pouvant nous aider à comprendre la masse de données qui nous entoure aujourd'hui et poser des fondations pour définir cette notion de savoir numérique. Ensuite, j'étudierai la nouvelle politique économique relative aux données qui s'est fait jour, comme moyen d'aborder quelques-unes des plus importantes relations de pouvoir qui sont en jeu dans notre nouvelle ère numérique.

19. Ackerman et Ball, art. cit.

1. Une vue socio-historique des *big data*

Je vais donc commencer par tracer les grandes lignes de la trajectoire historique du développement de ces mines de données, localiser quels secteurs ont favorisé leur accumulation et proposer différentes façons de catégoriser la masse d'informations que nous avons collectées jusqu'à aujourd'hui.

1.1. La trajectoire historique des *big data*

L'International Data Corporation (IDC), chercheur et consultant de premier ordre dans les technologies de l'information, fournit une mesure simple de l'étendue du phénomène en 2013 (IDC, 2014). Le montant global d'informations stockées dans ce que l'IDC appelle l'« univers numérique » est d'environ 4,4 zettaoctet (un zettaoctet est égal à 1,18059162 × 1021 octets). D'ici à 2020, ce nombre devrait augmenter pour atteindre les 44 zettaoctets (IDC, 2014).

La mesure la plus rigoureuse de la capacité technologique mondiale à stocker, à communiquer et à calculer des données a été présentée dans un article scientifique paru dans *Science* (Hilbert et López, 2011). (La plupart des descriptions de masses de données aujourd'hui et l'image folklorique de l'empilement de CD-ROM jusqu'à la Lune y ont puisé leur source.) Hilbert et López ont essayé de mesurer la capacité à réaliser ces trois tâches sur une période allant de 1986 à 2007, en comparant les technologies analogiques et numériques suivantes (fig. 1).

Technologie	**analogique**	**numérique**
Stockage	Vidéo analogique ; photo imprimée ; cassette audio ; photo négative ; ciné film ; disque LP ; épisode TV ; X Rays ; film TV ; journaux ; autres papiers et imprimés ; livres.	Disque-dur PC ; DVD et Blu-Ray; enregistrement numérique ; serveur et unité centrale ; CD et minidisque ; autres disques durs (portables) ; console ; carte mémoire ; téléphone mobile et PDA ; jeux vidéo ; caméras numériques et caméscope ; cartes à puce.
Communication	**Télédiffusion** : TV terrestre ; TV cable ; TV satellite ; radio ; journaux ; publicités en papier.	**Télédiffusion** : TV terrestre ; TV cable ; TV satellite ; radio ; navigation GPS.
	Télécommunication : téléphone fixe (voix) ; téléphone mobile (voix) ; lettres postales.	**Télécommunication** : téléphone fixe (voix) ; Internet ; téléphone mobile (voix) ; téléphone mobile (*data*).

Calcul		**Utilisation générale :** ordinateur ; jeux vidéo ; consoles ; serveur et unité centrale ; superordinateur ; calculatrice ; téléphone mobile/PDA.
		Processeurs de signaux numériques : CD, DVD et PVR, caméras et caméscope, modem et décodeur, GPS, média portable, imprimante et fax, radio, téléphone fixe et mobile; microcontroller ; processeur graphique.

Figure 1. Les technologies analogiques et numériques de stockage, de communication et de calcul.

Il est tout d'abord question de la capacité de stockage selon une perspective historique. La capacité globale n'a cessé de croître en raison du développement des technologies analogiques, au moins jusqu'en 2000. Mais après cela, la croissance a été exponentielle dans le domaine numérique, comme l'illustre la figure 2.

Technologie	1986	1993	2000	2007
analogique	2.62E+12	1.52E+13	4.08E+13	1.89E+13
numérique	2.08E+10	5.33E+11	1.37E+13	2.76E+14
Total	2.64E+12	1.58E+13	5.45E+13	2.95E+14

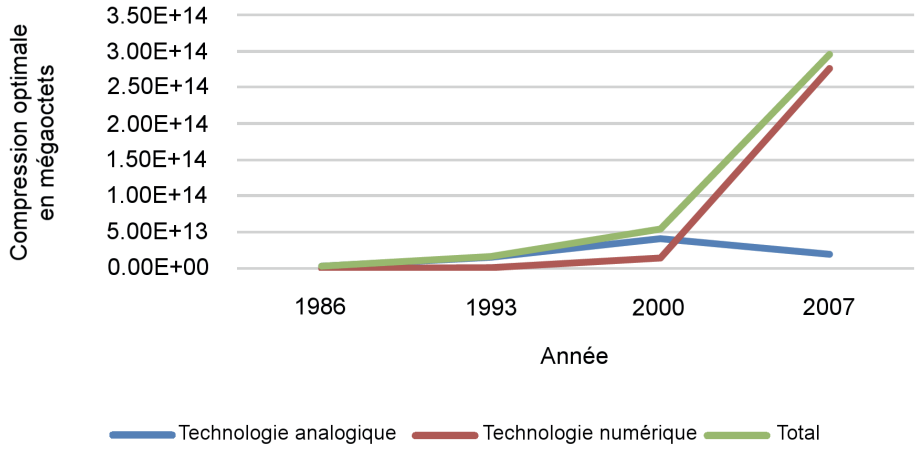

Figure 2. Capacité mondiale de stockage d'information en mégaoctets (Mo) selon le type de technologie.

Source : Hilbert et López, 2011, Table S A-1.

Ensuite, Hilbert et López mesurent la croissance de la capacité à diffuser et à communiquer l'information (fig. 3).

Technologie	1986	1993	2000	2007
Télédiffusion	4.32E+14	7.15E+14	1.15E+15	1.89E+15
analogique	4.32E+14	7.15E+14	1.07E+15	1.42E+15
numérique	-	-	8.37E+13	4.68E+14
Télécommunication	2.81E+11	4.71E+11	2.24E+12	6.54E+13
analogique	2.25E+11	1.48E+11	5.15E+10	3.63E+10
numérique	5.57E+10	3.23E+11	2.19E+12	6.53E+13
Total	4.32E+14	7.16E+14	1.15E+15	1.96E+15

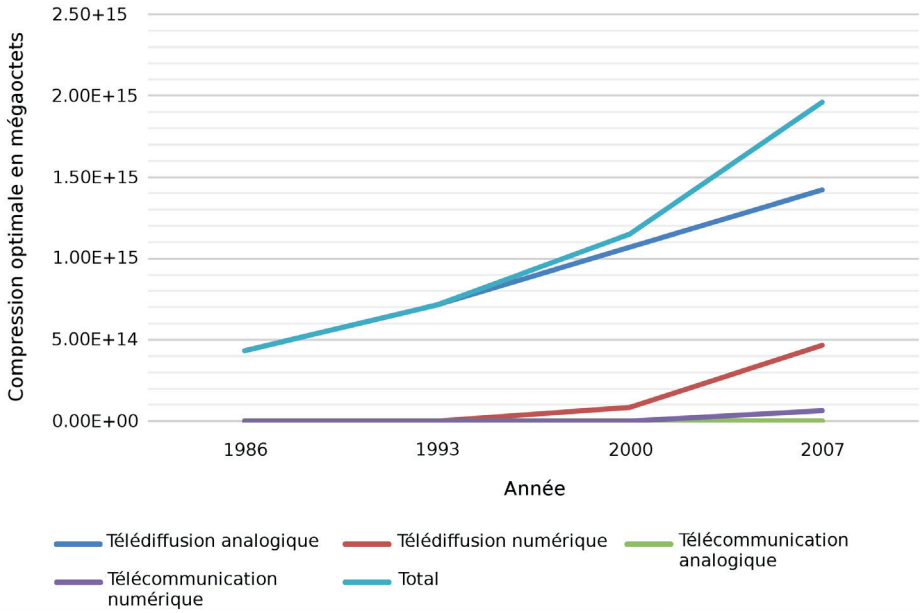

Figure 3. Capacité mondiale effective à diffuser et à communiquer l'information, en mégaoctets (Mo) selon le type de technologie et le secteur.

Source : Hilbert et López, 2011, pièces justificatives, p. 7.

Le calcul a quant à lui toujours été numérique, il n'y a donc pas lieu d'effectuer ici des comparaisons, mais soulignons l'augmentation exponentielle de la capacité technologique depuis 2000, comme cela est mis en évidence par la figure 4.

	1986	1993	2000	2007
Total	1.73E+07	4.02E+08	1.66E+10	5.38E+11

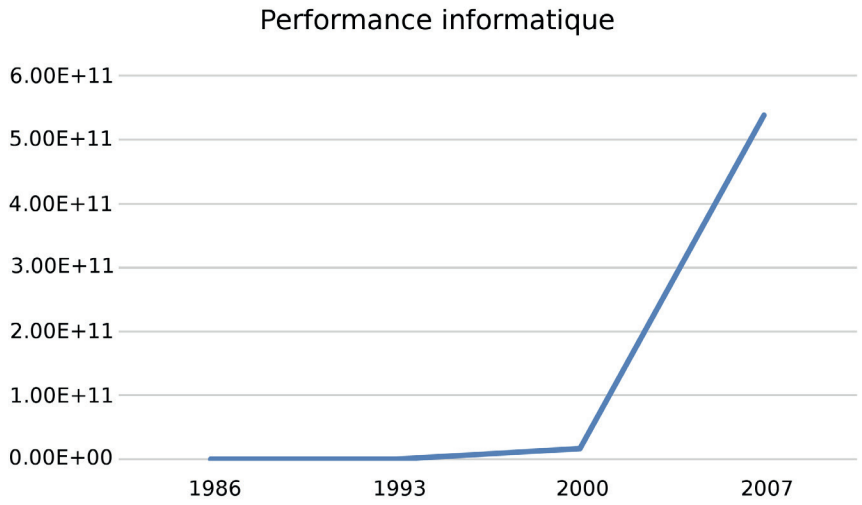

Figure 4. Capacité mondiale effective à traiter l'information sur des ordinateurs standard en millions d'instructions par seconde.

Source : Hilbert et López, 2011, pièces justificatives, p. 8.

L'étude de Hilbert et López révèle une augmentation remarquable dans le pourcentage annuel de croissance de la puissance informatique. Cela est représenté dans le graphique ci-dessous (fig. 5).

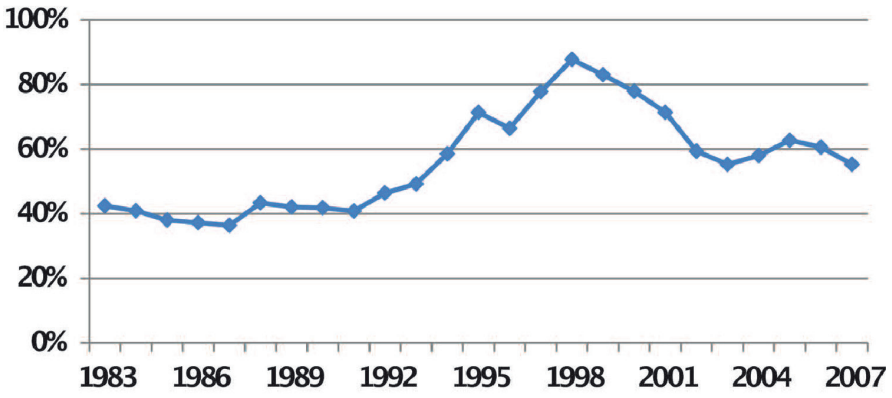

Figure 5. Croissance annuelle de la capacité de calcul comme pourcentage de l'ensemble des calculs antérieurs depuis 1977 (année *t* / S[1977, année *t* – 1]).

Source : Hilbert et López, 2011, p. 63.

Comme leurs données le démontrent, le pourcentage de capacité technologique repris par les médias numériques augmente à des taux remarquables, comme l'illustre le tableau ci-dessous (fig. 6).

		1986	1993	2000	2007	Taux annuel de croissance
Stockage numérique		0,8 %	3 %	25 %	94 %	23 %
Communication numérique	Télédiffusion	0 %	0 %	7,3 %	25 %	6 %
	Télé-communication	19,8 %	68,5 %	97,7 %	99,9 %	28 %
Calcul						58 %

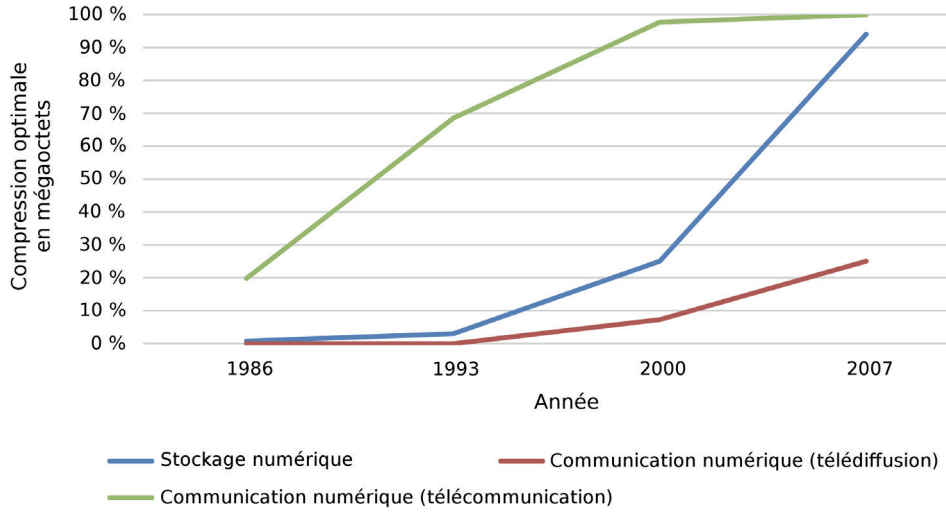

Figure 6. Évolution de la capacité mondiale de stockage et de communication d'informations selon le secteur (en pourcent).

Source : Hilbert et López, 2011, p. 63.

Nous savons, ensuite, que les *big data* ont augmenté exponentiellement récemment. La Commission fédérale des *big data* de la fondation TechAmerica estime que, en 2011, « 1,8 zettaoctets d'informations ont été créés, et ce montant devrait doubler tous les ans » (TechAmerica, 2012, p. 9). Ces projections sont cohérentes avec les chiffres du journal *The Economist*. Selon l'IDC, l'augmentation exponentielle devrait atteindre 35 zettaoctets en 2020 (TechAmerica, 2012, p. 11). Comme *The Economist* l'a souligné il y a plusieurs années, et comme cela est reflété dans les données précédentes, l'augmentation de la production des données surpasse largement notre capacité à les stocker. Le graphique ci-dessous (fig. 7) provient d'un rapport détaillé publié le 27 février 2010 – ce qui, dans notre ère numérique, semble être de la Préhistoire.

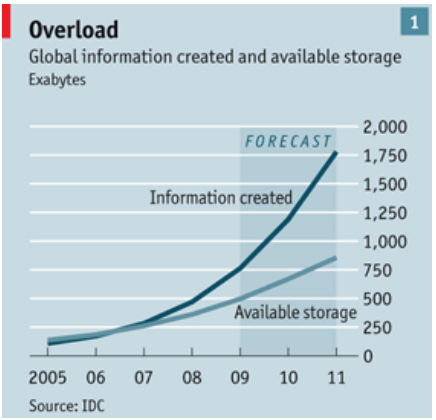

Figure 7. Surcharge informationnelle.

Prévision de la masse globale d'informations créées (courbe supérieure) et de la capacité de stockage (courbe inférieure) en exaoctets.

Source : « Data, data everywhere », *The Economist*, 25 février 2010 : http://www.economist.com/node/15557443.

McKinsey le souligne également : le taux de croissance des données dépasse largement notre capacité à les stocker (Manyika *et al.*, 2011, p. 16-17). C'est en 2007 que la création de données a excédé pour la première fois la capacité mondiale de stockage. Le total des données mondiales devrait augmenter de 40 % par an, la capacité de stockage et de calcul des données continuant également à augmenter – comme indiqué par Hilbert et López, le stockage ayant augmenté à un taux annuel de 23 % entre 1986 et 2007, tandis que la capacité de calcul augmentait à un taux de 58 % par an durant la même période.

En 2011, McKinsey a reporté quelques statistiques fascinantes à propos de ce qu'il a appelé ce « torrent grandissant » de données :

> 5 milliards de téléphones portables en marche en 2010 ;
> 30 milliards de contenus partagés sur Facebook par mois ;
> 235 téraoctet de données collectées par la bibliothèque du Congrès des États-Unis en avril 2011 ;
> 15 des 17 secteurs des États-Unis comptent plus de données stockées par société commerciale individuelle que la bibliothèque du Congrès des États-Unis.
>
> (Manyika *et al.*, 2011, préface)

Harvard Business Review rapporte qu'« à partir de 2012, environ 2,5 exaoctets de données sont créés chaque jour, et ce nombre double tous les 40 mois à peu près. Plus de données traversent l'Internet chaque seconde qu'il n'y en avait de stockées dans tout Internet il y a tout juste 20 ans ». À elle seule, l'entreprise Walmart « collecte plus de 2,5 pétaoctets de données chaque heure à partir des transactions de ses clients[20] ».

20. McAfee A. et Brynjolfsson E., « Big data: the management revolution », *Harvard Business Review*, octobre 2012, p. 62 : http://hbr.org/2012/10/big-data-the-management-revolution/ar.

1.2. La composition des *big data*

Les *big data* peuvent être présentées de manières différentes – selon des critères de forme, de source, de type, de but, etc. Voici quelques considérations utiles afin de décrire le torrent émergent de données.

1.2.1. Les sources de données

Selon McKinsey, en 2009, les trois premiers secteurs en termes de quantité d'informations stockées étaient :

1. *discrete manufacturing* (manufacture et fabrication de pièces détachées) (966 pétaoctets) ;
2. le gouvernement (848 pétaoctets) ;
3. la communication et les médias (715 pétaoctets) (Manyika *et al.*, 2011, p. 19).

L'IDC rapporte en 2013 qu'environ deux tiers de l'information dans l'univers numérique, soit environ 2,9 zettaoctets, sont « générés par les consommateurs », le reste, soit environ 1,5 zettaoctets, étant généré par les entreprises. Le monde des affaires touche cependant environ 85 %, soit 2,3 zettaoctets, des données, contrairement aux 15 %, représentant environ 0,6 zettaoctets, qui ne sont pas touchés par les entreprises (Turner *et al.*, 2014).

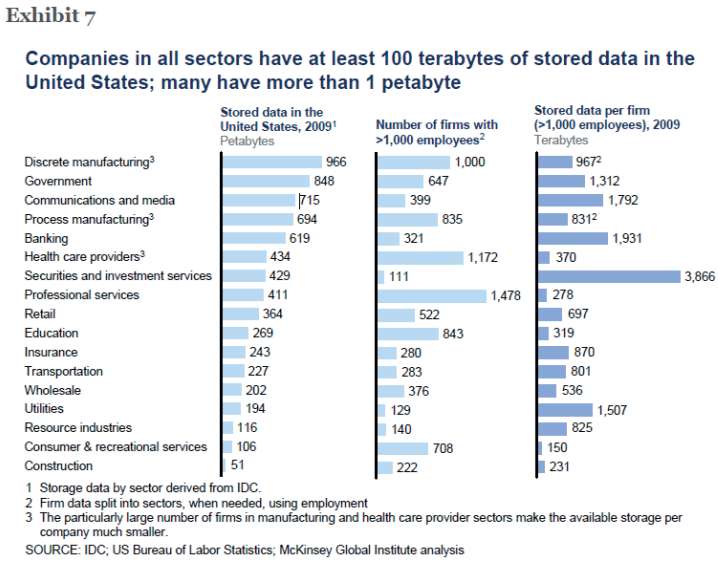

Figure 8. Stockage des données aux États-Unis et dans les entreprises américaines.
Source : IDC ; US Bureau of Labor Statistics, McKinsey Global Institute Analysis.

Seule une partie des données est générée ou collectée par les gouvernements. Malgré cela, il est exact que, depuis 2000, aux États-Unis, « le montant d'informations dont le gouvernement fédéral s'empare a augmenté exponentiellement » (TechAmerica, 2012, p. 9). Pour en donner une idée, en 2009 uniquement, « le gouvernement des États-Unis a produit 848 pétaoctets de données et le système de sécurité sociale américain a, à lui seul, atteint 150 exaoctets. Cinq exaoctets (10^18 gigaoctets) de données suffiraient pour contenir tous les mots qui ont été prononcés par des êtres humains sur la Terre. À cette allure, les *big data* du système de sécurité sociale américain atteindront bientôt le zétaoctet (10^21 gigaoctets), puis le yottaoctet (10^24 gigaoctets) » (TechAmerica, 2012, p. 9).

Les autres secteurs de l'économie qui stockent la plupart des données du pays peuvent être visualisés sur le graphique auparavant (fig. 8), préparé par McKinsey (Manyika *et al.*, 2011, p. 19).

1.2.2. Les types de données I : intelligentes, d'identité et de la personne

Higinio Maycotte, dans un article publié dans *Wired*, suggère qu'il y a trois types particuliers de *big data* qui devraient être distingués : les « données intelligentes », les « données d'identité » et les « données de la personne[21] ».

Les « données intelligentes » désignent la partie des *big data* qui a été « cloisonnée, segmentée, puis visualisée » pour un besoin particulier[22]. Ces données ont été rendues lisibles à partir de leur état original de collection de nombres binaires, et ne requièrent plus l'intervention d'un expert pour être analysées.

Les « données d'identité » sont peut-être le type le plus important de *big data*. Elles constituent « la force derrière la modélisation prédictive et l'apprentissage par la machine », et visent à « raconter l'histoire de qui vous êtes dans l'ère numérique, en incluant ce que vous aimez, ce que vous achetez, vos choix relatifs à votre style de vie et quand et à quelles intervalles tout cela arrive[23] ». Cela inclut les médias sociaux, les achats, les analyses comportementales, et d'autres choses encore. Comme Maycotte le fait remarquer, lorsque les fichiers clients de Target ont été piratés l'année dernière, « c'était la perte des données d'identité […] qui était devenu le plus gros problème » ; dans cette affaire, l'identité volée correspondait aux « numéros de cartes bancaires associés aux noms et aux adresses physiques, de même que les adresses mail[24] ».

Enfin, les « données de la personne » sont créées par l'agrégation de données sociales à travers le temps[25]. Elles sont collectées en regardant les « données échappées » d'un grand nombre d'utilisateurs : « qui est-ce que votre public aime et suit sur les réseaux sociaux, sur quels liens ils cliquent, combien de temps ils restent sur le site sur lequel ils ont cliqué et combien de convertis et de non-convertis[26] ».

---

21. Maycotte H., « The evolution of big data, and where we're headed », *Wired*, 26 mai 2014 : http://www.wired.com/2014/03/evolution-big-data-headed.
22. *Ibid.*
23. *Ibid.*
24. *Ibid.*
25. *Ibid.*
26. *Ibid.*

1.2.3. Les types de données II : sociales, d'appareils et transactionnelles

Cette première division tripartite recoupe, en partie, une autre façon de disséquer les données numériques en « données sociales », « données d'appareils » et « données transactionnelles[27] ».

Les données sociales sont en grande partie assimilables aux données de la personne décrites précédemment par Maycotte. Elles sont majoritairement générées par le comportement du consommateur sur l'Internet. Cela inclut les données des médias sociaux, qui sont massives : on compte ainsi « 230 millions de tweets postés sur Twitter par jour, 2,7 milliards de "J'aime" et de commentaires ajoutés sur Facebook chaque jour, et 60 heures de vidéos téléchargées vers le serveur de YouTube chaque minute[28] ».

Les données d'appareils sont des « informations générées par l'équipement industriel, les données en temps réel fournies par les capteurs qui surveillent et contrôlent les pièces des machines (souvent aussi appelé l'« Internet des objets »), et même les blogs en ligne qui suivent le comportement de l'utilisateur en ligne[29] ». Une illustration de cela réside dans le Large Hadron Collider (LHC) au CERN, le plus grand centre de recherche sur les particules physiques au monde, qui génère approximativement 40 téraoctets de données par seconde lorsque des expériences sont en cours.

Enfin, les données transactionnelles sont générées par l'agrégation et l'enregistrement des transactions quotidiennes d'une entreprise, en incluant les articles vendus, « la carte d'identité des produits, les prix, les informations relatives au paiement et les données du fabricant et du distributeur[30] ». Pensez ici à Amazon ou à Domino's Pizza – ce dernier sert environ un million de consommateurs par jour – qui produisent d'énormes quantités de *big data* de manière journalière.

1.2.4. Les types de données III : des individus, du secteur public et du secteur privé

Le Forum économique mondial (FEM) souligne que les données sont générées à partir de trois secteurs différents – ce qui nous donne une autre répartition tripartite. Le premier secteur est celui des individus, qui fournissent un type particulier de données : données de l'« externalisation ouverte », données sociales des téléphones portables, et d'autres « échappements de données ». Le FEM souligne, à cet égard, que « les données émanant des téléphones portables sont particulièrement prometteuses, en partie parce que pour beaucoup de personnes à faibles revenus, c'est la seule forme de technologie interactive, mais aussi parce qu'il est plus facile de lier les données générées par les mobiles aux individus » (Forum économique mondial, 2012, p. 2). Et des tonnes de données de ce type sont en train d'être produites : « les transactions

27. « Big data FAQs – A primer », *Arcplan Business Intelligence Blog*, 23 mars 2012 : http://www.arcplan.com/en/blog/2012/03/big-data-faqs-a-primer.
28. *Ibid*.
29. *Ibid*.
30. *Ibid*.

financières en ligne ou par mobile, le trafic des réseaux sociaux, et les coordonnées des GPS génèrent maintenant plus de 2,5 quintillions d'octets de prétendues *big data* tous les jours[31] ».

Le deuxième secteur comprend le secteur public et celui du développement. Cela inclut notamment les données de recensement du gouvernement, les statistiques et les indicateurs de santé publique, les données relatives aux taxes et aux dépenses, et d'autres « données des installations ».

Le troisième secteur est le secteur privé, et il comprend les données des transactions, des achats, des dépenses, de la consommation et utilise les informations[32].

Telles sont quelques-unes des façons de catégoriser le torrent de données qui est en train d'émerger dans cette ère numérique – et qui peuvent nous aider à comprendre ce nouvel univers numérique.

2. Une économie politique des données

Avec la croissance exponentielle de la capacité technologique, on a vu émerger une nouvelle et véritable politique économique des données. En 2012 uniquement, l'industrie du courtage en données a atteint un revenu de 156 milliards de dollars, ce qui, comme l'a souligné le sénateur John D. Rockefeller IV, représente « deux fois la taille du budget global des services de renseignement du gouvernement des États-Unis – tout cela rendu possible par l'effort de détailler et de vendre les informations relatives à nos vies privées[33] » (U.S. Senate Committee on Commerce, Science, and Transportation, 2013). Il y a, aujourd'hui, plus de 4000 entreprises de courtage en données, certaines d'entre elles cotées en bourse ou portant des noms connus de tout le monde, comme Lexis-Nexis ou Experian, et beaucoup d'autres plus petites et moins bien connues[34].

2.1. Le marché des données

Le sénateur Rockefeller a tenu des audiences au sein du comité du Sénat des États-Unis sur le Commerce, la Science et le Transport le mercredi 18 décembre 2013, afin d'examiner cette industrie du courtage en données – et de mettre en lumière certaines pratiques assez controversées[35]. Ces audiences ont révélé, par exemple, qu'un courtier en données à Lake Forest, dans l'Illinois, Medbase200, a proposé de vendre à des compagnies pharmaceutiques

31. World Economic Forum (2012), *Big Data, Big Impact: New Possibilities for International Development* : http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf, p. 2.
32. *Ibid.*, p. 5.
33. Cf. aussi Levine Y., « You'll be shocked at what "surveillance valley" knows about you: The most intimate details about a person's life, packaged and sold over and over again to anyone willing to pay », *AlterNet.org*, 8 janvier 2014 : http://www.alternet.org/media/what-surveillance-valley-knows-about-you.
34. Levine, art. cit.
35. Pasquale F., « The dark market for personal data », *New York Times*, 17 octobre 2014 : http://www.nytimes.com/2014/10/17/opinion/the-dark-market-for-personal-data.html.

une liste de « souffrants de viol » au prix de 79 dollars pour 1 000 noms[36]. Medbase200 vendait cette liste sur son site Internet de la manière suivante :

> Ces souffrants de viol sont des membres de familles qui ont dit, ou ont été identifiés comme souffrant de maladies, de conditions ou d'affections spécifiques associées au viol[37]…

Medbase200 a retiré la base de données des « souffrants de viol » de son site internet après ces révélations, de même que les « listes de victimes de violence domestique, des patients atteints du sida et des malades de la pression exercée par l'entourage qui étaient offertes à la vente[38] ». Mais le nombre et la variété d'autres listes à vendre sont tout simplement stupéfiants. Voici les listes à vendre commençant par la lettre « A ». La taille de chaque base de données et les informations sur le prix (en dollars pour 1 000 renseignements) figurent dans la colonne de droite (fig. 9).

Allergy/Immunology Nurses	53423 Total Universe @ 59/M
AARP Members Mailing List	20435556 Total Universe @ 79/M
Abscess Sufferers	> (Inquire) Total Universe @ 79.00/M
Abuse Sufferers	> (Inquire) Total Universe @ 79.00/M
Acetaminophen Users	21092445 Total Universe @ 79/M
Achondroplasia Sufferers	> (Inquire) Total Universe @ 79.00/M
Acid Reflux Disease (GERD) Sufferers	> (Inquire) Total Universe @ 79.00/M
Acid Reflux Disease (GERD) Sufferers at Home	5456709 Total Universe @ 79/M
Acne Sufferers	> (Inquire) Total Universe @ 79.00/M
Addiction Sufferers	> (Inquire) Total Universe @ 79.00/M
Addiction/Substance Abuse (Drug Abuse) Nurses	38009 Total Universe @ 59/M
Addison's Disease Sufferers	> (Inquire) Total Universe @ 79.00/M
Adenoma Sufferers	> (Inquire) Total Universe @ 79.00/M
Adolescent Medicine Nurses	20198 Total Universe @ 59/M
Adult Medicine/Adult Care Nurses	98996 Total Universe @ 59/M
Advanced Practice Nurses	92231 Total Universe @ 59/M
Aestheticians at Home	116545 Total Universe @ 59/M
Agoraphobia Sufferers	> (Inquire) Total Universe @ 79.00/M
Aids And Hiv Infection Sufferers	> (Inquire) Total Universe @ 79.00/M
AIDS/HIV Nurses	300893 Total Universe @ 59/M
Ailments, Diseases & Conditions - Hispanic Sufferers	17234554 Total Universe @ 79/M

36. Dwoskin E., « Data broker removes rape-victims list after journal inquiry », *Wall Street Journal Digits*, 19 décembre 2013 : http://blogs.wsj.com/digits/2013/12/19/data-broker-removes-rape-victims-list-after-journal-inquiry.
37. Une copie de la liste est accessible ici : https://www.evernote.com/shard/s1/sh/235f0aab-785b-4a10-bf3b-f39b3fd0dec7/ed697225af44a19f18240183df03cd0f.
38. Dwoskin E., « Data broker removes rape-victims list after journal inquiry », *WSJ Blogs*, 19 décembre 2013 : http://blogs.wsj.com/digits/2013/12/19/data-broker-removes-rape-victims-list-after-journal-inquiry.

Ailments, Diseases & Conditions - Sufferers	227453121 Total Universe @ 79/M
Ailments, Diseases & Conditions - Sufferers (Vol)	227453121 Total Universe @ 39.5/M
Ailments, Diseases & Conditions - Sufferers via Email	173209889 Total Universe @ 129/M
Albinism Sufferers	> (Inquire) Total Universe @ 79.00/M
Alcoholic Hepatitis Sufferers	> (Inquire) Total Universe @ 79.00/M
Alcoholism Sufferers	> (Inquire) Total Universe @ 79.00/M
Allergies Sufferers	> (Inquire) Total Universe @ 79.00/M
Allergy / Immunology Nurses	57886 Total Universe @ 59/M
Allergy Sufferers at Home	25698121 Total Universe @ 79/M
Alli Users	1985452 Total Universe @ 79/M
Alopecia (Thinning Hair/Hair Loss) Sufferers	> (Inquire) Total Universe @ 79.00/M
Altitude Sickness Sufferers	> (Inquire) Total Universe @ 79.00/M
Alzheimer's Disease Sufferers	> (Inquire) Total Universe @ 79.00/M
Amblyopia Sufferers	> (Inquire) Total Universe @ 79.00/M
Ambulatory Care Nurses	72234 Total Universe @ 59/M
Amebiasis Sufferers	> (Inquire) Total Universe @ 79.00/M
Amnesia Sufferers	> (Inquire) Total Universe @ 79.00/M
Amyotrophic Lateral Sclerosis Sufferers	> (Inquire) Total Universe @ 79.00/M
Anemia Sufferers	> (Inquire) Total Universe @ 79.00/M
Anesthesiology Nurses	172339 Total Universe @ 59/M
Aneurdu Sufferers	> (Inquire) Total Universe @ 79.00/M
Aneurysm Sufferers	> (Inquire) Total Universe @ 79.00/M
Angina Sufferers	> (Inquire) Total Universe @ 79.00/M
Animal Bites Sufferers	> (Inquire) Total Universe @ 79.00/M
Anorexia Sufferers	> (Inquire) Total Universe @ 79.00/M
Anosmia Sufferers	> (Inquire) Total Universe @ 79.00/M
Anotia Sufferers	> (Inquire) Total Universe @ 79.00/M
Anthrax Sufferers	> (Inquire) Total Universe @ 79.00/M
Antisocial Personality Disorder Sufferers	> (Inquire) Total Universe @ 79.00/M
Anxiety And Anxiety Disorders Sufferers	> (Inquire) Total Universe @ 79.00/M
Anxiety Disorders Sufferers	> (Inquire) Total Universe @ 79.00/M
Anxiety Sufferers (GAD) Sufferers at Home	3983434 Total Universe @ 79/M
Appendicitis Sufferers	> (Inquire) Total Universe @ 79.00/M
Apraxia Sufferers	> (Inquire) Total Universe @ 79.00/M
Argyria Sufferers	> (Inquire) Total Universe @ 79.00/M
Arthritis Nurses	180371 Total Universe @ 59/M
Arthritis Sufferers	> (Inquire) Total Universe @ 79.00/M
Arthritis, Infectious Sufferers	> (Inquire) Total Universe @ 79.00/M
Ascariasis Sufferers	> (Inquire) Total Universe @ 79.00/M

Aseptic Meningitis Sufferers	> (Inquire) Total Universe @ 79.00/M
Asperger Disorder Sufferers	> (Inquire) Total Universe @ 79.00/M
Asthenia Sufferers	> (Inquire) Total Universe @ 79.00/M
Asthma Sufferers	> (Inquire) Total Universe @ 79.00/M
Astigmatism Sufferers	> (Inquire) Total Universe @ 79.00/M
Atherosclerosis Sufferers	> (Inquire) Total Universe @ 79.00/M
Athetosis Sufferers	> (Inquire) Total Universe @ 79.00/M
Athlete's Foot Sufferers	> (Inquire) Total Universe @ 79.00/M
Atrophy Sufferers	> (Inquire) Total Universe @ 79.00/M
Attention Deficit Hyperactivity Disorder (ADHD) Sufferers	> (Inquire) Total Universe @ 79.00/M
Attention Sufferers	> (Inquire) Total Universe @ 79.00/M
Autism Sufferers	> (Inquire) Total Universe @ 79.00/M
Autism Sufferers at Home	2983342 Total Universe @ 79/M
Avandia Users	6898545 Total Universe @ 79/M

Figure 9. Listes à vendre commençant par la lettre « A ».

Source : Medbase200 I Medical Marketing Lists (Dec. 18, 2013) : https://www.evernote.com/shard/s1/sh/b263bb78-c5f0-404e-9de7-ba86ee3b9369/5f3f08f3cc32e676adcc523e62e1422d.

Comme *The New York Times* l'a récemment rapporté, InfoUSA, un des plus importants courtiers en données aux États-Unis, « a effectué de la publicité pour des listes de "personnes âgées à la recherche d'opportunités", 3,3 millions de personnes plus âgées "cherchant des moyens de se faire de l'argent", et les "seniors souffrants", 4,7 millions de personnes atteintes d'un cancer ou de la maladie d'Alzheimer. "Vieillots mais Cadeaux" contenait 500 000 parieurs de plus de 55 ans, pour 8,5 centimes chacun. Une des listes disait : "Ces gens sont crédules. Ils veulent croire que la roue peut tourner[39]." » Comme vous pouvez l'imaginer, ces types de listes sont souvent vendues à des gens sans scrupules qui vont ensuite s'attaquer aux personnes listées[40].

Un des courtiers en données est appelé l'Acxiom Corporation et vu, selon un autre journaliste du *New York Times*, comme « le géant silencieux d'une industrie à multimilliard de dollars connue sous le nom de *database marketing*[41] ». « Peu de consommateurs ont déjà entendu parler de Acxiom », rapporte *The Times*. « Mais les analystes disent qu'il a amassé la plus grosse base de données mondiale sur les consommateurs – et qu'il veut en savoir plus, beaucoup plus. Ses serveurs traitent plus de 50 trillions de "transactions" de données par an. Les dirigeants de l'entreprise ont dit que sa base de données contient des informations sur environ 500 millions de consommateurs actifs à travers le monde, avec environ 1 500 points d'information par personne. Cela inclut une majorité des adultes aux États-Unis[42] ».

39. Duhigg C., « Bilking the elderly, with a corporate assist », *The New York Times*, 20 mai 2007 : http://www.nytimes.com/2007/05/20/business/20tele.html ; cf. aussi (Pasquale, 2015).
40. *Ibid*.
41. Singer N., « Mapping, and sharing, the consumer genome », *The New York Times*, 16 juin 2012 : http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html
42. *Ibid*.

Les données de consommation sont devenues une marchandise, son commerce un marché hautement compétitif[43]. Le taux en vigueur pour l'information sur les consommateurs individuels varie en fonction de son caractère, de la compétition entre les fournisseurs de données et de la « pure ubiquité de détails à propos de centaines de millions de consommateurs[44] ». En général, cependant, « plus l'information est intime, plus elle a de valeur ». Par exemple, Steel écrit qu'« une information basique telle que l'âge, le sexe et la localisation se vend pour 0,0005 dollars par personne, ou 0,50 dollar par centaine de personnes, selon les détails des prix vus par le *Financial Times* ». Mais l'information à propos d'individus « soupçonnés d'être "influents" dans leurs réseaux sociaux se vend pour 0,00075 dollar, ou 0,75 par centaine de personnes ». Les données à propos de « revenus et d'historique des achats » se vendent pour 0,001 dollar – donc, un dollar par mille[45].

Les données sur les consommateurs sont utilisées pour prédire un futur comportement d'achat, pour classer les comportements existant dans des « catégories prédéterminées » ; pour associer certains comportements entre eux, comme lorsque, pour emprunter un exemple d'un article récent paru dans *The Atlantic*, Amazon fait une recommandation pour des verres à Martini fondée sur l'achat récent d'un *shaker* ; et pour former des « groupes » d'informations fondés sur les comportements consuméristes, comme lorsqu'un groupe de consommateurs est distingué selon ses loisirs et ses intérêts[46]. La surveillance du consommateur, comme le note Lyon (2003, p. 172), est « la sphère de surveillance qui grandit le plus rapidement […], dépassant les capacités de surveillance de la plupart des États. Et même au sein des États, la surveillance administrative est guidée autant par les canons de la consommation que par ceux de la citoyenneté, interprétés de façon classique ».

2.2. Le libre marché de la surveillance

Qu'est-ce qui est vraiment libre dans le libre marché des données ? Peu de choses, excepté, peut-être, l'accès du gouvernement à celles-ci. Une chose qui semble ne rien coûter, ou presque, c'est la possibilité pour le gouvernement d'accéder à ces mines d'informations. C'est ce qui, en partie, est en train de déconstruire le lien entre l'État et l'échange.

Le programme PRISM en est un bon exemple. Lancé en 2007, il permet à la NSA d'accéder aux données de Microsoft, Yahoo, Google, Facebook, PalTalk, YouTube, Skype, AOL, Apple, et d'autres – pour un coût total du programme de seulement 20 millions de dollars par an[47].

---

43. *Ibid.*

44. Steel E., « Financial worth of data comes in at under a penny a piece », *Financial Times*, 12 juin 2013 : http://search.proquest.com/docview/1367072051.

45. Steel E., « Companies in scramble for consumer data », *Financial Times*, 12 juin 2013. Cf. aussi Steel E., Locke C., Cadman E. et Freese B., « How much is your personal data worth? », *Financial Times*, 12 juin 2013 : http://www.ft.com/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html.

46. Furnas A., « Everything you wanted to know about data mining but were afraid to ask », *The Atlantic*, 3 avril 2012 : http://www.theatlantic.com/technology/archive/2012/04/everything-you-wanted-to-know-about-data-mining-but-were-afraid-to-ask/255388.

47. Greenwald G. et MacAskill E., « NSA PRISM program taps in to user data of Apple, Google and others », *The Guardian*, 6 juin 2013, accessible http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data.

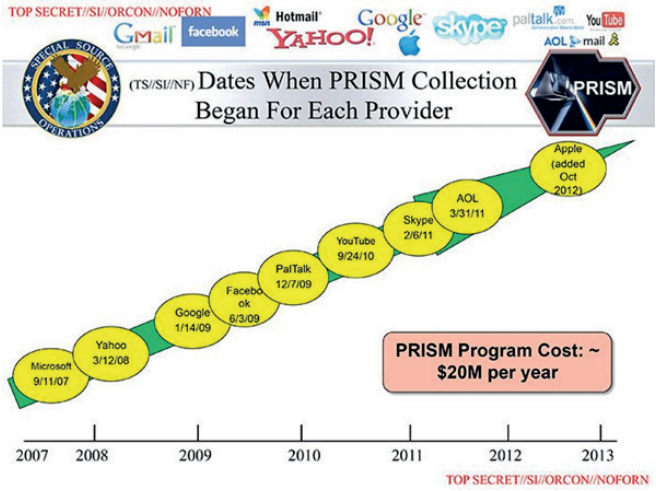

Figure 10. Diapositive PowerPoint utilisée par la NSA pour décrire l'histoire du programme PRISM, telle qu'elle a été divulguée par Edward Snowden et publiée par le *Guardian*.

Le programme PRISM donne accès au gouvernement aux mails des individus, à leurs photos, vidéos, pièce jointes, VoIP (*Voice over Internet Protocols*), etc. À un coût pratiquement nul pour le gouvernement, il fournit un accès presque complet à leurs vies numériques. Et, bien que la NSA dénie qu'elle ait un accès immédiat aux données, les documents de la NSA divulgués par Snowden établissent que l'agence « revendique "la collection directement depuis les serveurs" des fournisseurs majeurs des États-Unis[48] ». Le type de données disponibles est large : comme le *Guardian* le rapporte, « le programme PRISM permet aux services de renseignement d'accéder directement aux serveurs des compagnies. Le document de la NSA note que les opérations ont l'"assistance des fournisseurs de communications des États-Unis[49]" ».

Selon le *Guardian*, il n'est pas nécessaire d'obtenir un mandat ou une autorisation individuelle sous le Foreign Intelligence Surveillance Act (FISA) pour ces activités de collecte et d'analyse tant que l'analyste cherchant les communications a « un doute raisonnable que l'une des parties était en dehors du pays au moment où les enregistrements ont été collectés par la NSA[50] ».

Le programme est apparemment en train de conduire à une augmentation exponentielle des requêtes de recherche.

> Le document met en lumière l'augmentation du nombre de communications obtenues en 2012 de 248 % pour Skype – conduisant les notes à remarquer qu'il y avait une "croissance exponentielle dans les signalements Skype ; il semble que notre aptitude contre Skype commence à être connue" » rapporte le *Guardian*. « Il y eu également une augmentation de 131 % des requêtes pour les données Facebook, et de 63 % pour Google[51].

48. *Ibid.*
49. *Ibid.*
50. *Ibid.*
51. *Ibid.*

Conclusion

L'ère du spectacle dans l'Antiquité était une preuve du coût de la publicité. Rendre quelque chose public était coûteux, et les Anciens devaient donc s'amasser dans un lieu commun pour observer, partager, prendre part ensemble à l'acte de voir. L'ère de la surveillance au XIX^e^ siècle, en revanche, nous a offert une preuve du coût de la sécurité. Surveiller et corriger les individus était coûteux. L'invention des nouvelles institutions de détention a été conduite, en grande partie, par la recherche de moyens plus efficients.

Nous sommes maintenant entrés dans une ère de publicité sans coût. Aujourd'hui, cela ne coûte rien de diffuser son information. Au contraire, l'information elle-même a de la valeur. La plupart des personnes l'abandonnent gratuitement. Dans beaucoup de cas, on les y a forcés – nous sommes devenus nos propres administrateurs et passons notre temps à faire toute notre propre « administration ». Nous entrons nos propres données personnelles sur des sites internet de réservation de billets d'avion, sur Zappos.com, sur Amazon. Et c'est précisément cette ressource naturelle qui a donné lieu à une nouvelle politique économique des données – une politique économique de la publication, de la publicité, du « rendre public ».

Qu'advient-il lorsque nos modes de communication eux-mêmes et nos actes d'existence de tous les jours – quand chaque mot que nous prononçons, chaque note prise, chaque lettre écrite, toute photo et toute vidéo, même notre battement de cœur, nos taux de cholestérol, nos caractéristiques faciales, nos retraits bancaires, nos validations de tickets de métro, etc. – peuvent être enregistrés, stockés, reliés, extraits et analysés par des machines avec des capacités de calcul qui excèdent de loin celles du cerveau humain ? La réponse, semble-t-il, est un marché des données florissant qui rend possibles des formes de contrôle et de surveillance que nous n'aurions jamais imaginées – excepté, peut-être, dans les dystopies orwelliennes.

Cette nouvelle et émergente politique économique des données a été rendue possible non seulement par l'innovation technologique, mais également par une nouvelle forme de pouvoir qui nous a enseigné à céder volontairement nos informations lorsqu'elles étaient demandées, à nous identifier, à révéler nos plus profonds secrets, à nous conformer aux requêtes – et ironiquement, dans un monde de propriété privée, à ne jamais sentir que nous avions droit à une propriété privée sur notre propre identité et toutes ces informations personnelles.

Les conséquences sont considérables, en particulier dans la mesure où elles brisent les frontières traditionnelles qui séparent gouverner d'échanger et de sécuriser.

Références

Hilbert M. et López P. (2011), « The world's technological capacity to store, communicate, and compute information », *Science*, vol. 332, nº 6025, p. 60-65, DOI : 10.1126/science.1200970.

IDC (2014), *The Digital Universe of Opportunities: Rich Data & the Increasing Value of the Internet of Things* : https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf.

Lyon D. (2003), « Surveillance technology and surveillance society », *in* Misa T., Brey P., et Feenberg A. (dir.), *Modernity and Technology*, Cambridge, Mass., MIT Press, p. 161-183.

Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C. et Hung Byer A. (2011), *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey and Company : http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Pasquale F. (2015), *The Black Box Society*, Cambridge (Mass.), Harvard University Press.

TechAmerica (2012), *Demystifying Big Data: A Practical Guide to Transforming the Business of Government*, Report, TechAmerica Foundation's Federal Big Data Commission : http://www.techamerica.org/Docs/techAmerica-BigDataReport-FINAL.pdf.

Turner V., Gantz J. F., Reinsel D. et Minton S. (2014), *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*, IDC : http://idcdocserv.com/1678.

U.S. Senate Committee on Commerce, Science, and Transportation (2013), Statement of Senator John D. (Jay) Rockefeller IV at the Senate Hearings : « What information do data brokers have on consumers, and how do they use it? », 18 décembre 2013.

World Economic Forum (2012), *Big Data, Big Impact: New Possibilities for International Development* : http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf.

La contribution des internautes aux *big data* : un travail ?

Guillaume Tiffon
Maître de conférences en sociologie à l'université d'Évry et chercheur au centre Pierre-Naville

Sur Internet, on le sait, chaque navigation, chaque vidéo, chaque photographie et chaque commentaire postés laissent une trace ; une trace numérique que certaines entreprises collectent, puis revendent. En cela, les internautes, qu'ils soient des consommateurs ou de simples utilisateurs, contribuent, le plus souvent sans le vouloir ni même le savoir, à la production de données massives, aussi appelées *big data*. Loin d'être marginale, cette marchandisation des traces numériques se situe au cœur du modèle économique de certains géants de la « Net économie », comme Google ou Facebook, et pose question quant au rôle des internautes dans ces *success stories* : sont-ils mis au travail ? Leur contribution crée-t-elle de la valeur, que ces entreprises s'accaparent ? Ces empires bâtis en moins de deux décennies reposent-ils sur un mécanisme d'extraction de valeur, un renouvellement des formes contemporaines d'exploitation ? Autant de questions auxquelles ce chapitre propose de répondre.

Après une brève présentation des travaux sur la mise au travail des clients, je reviendrai sur le caractère polysémique de la notion de travail, pour expliciter à partir de quelle acception et sous quelles conditions l'activité des clients peut être qualifiée comme telle. Le cas plus spécifique des traces numériques, et de la nature du travail qu'elles nécessitent, sera ensuite abordé ; ce qui me permettra, pour finir, de revenir sur les questions que soulèvent l'utilisation et la marchandisation de ces grandes bases de données numériques.

1. Les modèles d'analyse de la mise au travail des clients

Il existe dans la littérature deux manières d'appréhender la question de la mise au travail des clients. L'une consiste à se placer du point de vue de l'entreprise ; l'autre, du travail et des travailleurs, en resituant l'activité déployée au sein des rapports de production capitalistes. Je présenterai chacune d'entre elles avant d'exposer plus précisément notre modèle d'analyse.

1.1. Le marketing des services

Historiquement, la mise au travail des clients a d'abord été étudiée en marketing des services, où, dès les années 1980, la coproduction est présentée comme une opportunité dont les entreprises doivent se saisir pour accroître leurs performances (Eiglier et Langeard, 1987).

Dans ces travaux, les spécialistes du marketing conseillent aux entreprises de « reporter sur le client une partie du travail auparavant effectué par le personnel » (Eiglier, 2004, p. 39). Cela permettrait d'augmenter sensiblement la productivité dans les services et constituerait un atout compétitif considérable. Pour Neeli Bendapudi et Robert P. Leone (2003), ce serait même la prochaine limite de la compétitivité. Pour ce faire, ces spécialistes recommandent aux entreprises de manager leurs clients comme des « employés partiels » (Lovelock et Young, 1979 ; Mills et Morris, 1986 ; Kelley, Donnely et Skinner, 2001) et d'adopter à leur égard les mêmes principes de gestion des ressources humaines que pour les salariés (Bowen, 1986 ; Larsson et Bowen, 1989 ; Chase et Dasu, 2001). Ainsi, ces dernières devraient établir des procédures et des descriptions de poste pour contrôler et évaluer le travail des clients (Mills, Chase et Margulies, 1983), ne pas hésiter à les former afin de mieux exploiter leurs compétences et leurs talents (Prahalad et Ramaswamy, 2000), et envisager, s'il le faut, de mettre en place une politique de recrutement des clients, selon leurs performances de coproducteurs (Lapert, 2005). Dans ces analyses, les auteurs ont ceci en commun d'adopter le point de vue de l'entreprise et de se poser d'emblée en prescripteurs de solutions pour améliorer la performance des firmes.

1.2. La sociologie du travail des clients

En sociologie, l'analyse du phénomène est plus récente. Elle apparaît en France à la fin de la première décennie des années 2000 (Dujarier, 2008 ; Tiffon, 2013a ; Bernard, Dujarier et Tiffon, 2011a et 2011b) et s'inscrit dans une perspective quelque peu différente. Dans le sillage de la sociologie du travail classique, l'activité des travailleurs, qu'ils soient salariés ou consommateurs, y est analysée comme un acte technique, dont les conditions de réalisation comme les finalités sont le produit de rapports de production historiquement spécifiques, qu'il convient de déconstruire et de questionner. En cela, la sociologie du travail des clients se distingue nettement du marketing des services, qui, par sa posture, naturalise de fait le mode de coordination par le marché et ses effets.

Il existe néanmoins des différences internes à la sociologie du travail des clients. En France, on peut identifier deux modèles, qui se complètent plus qu'ils ne s'opposent : l'un est formalisé par Marie-Anne Dujarier (2008) ; l'autre par moi-même, dans des articles éparses, à partir de 2007, puis dans un ouvrage synthétisant mes réflexions sur le sujet (Tiffon, 2013a). Partant de situations quotidiennes au sein de services marchands, le premier modèle identifie trois formes de mise au travail des clients :

- l'externalisation de tâches simplifiées, contraignant les consommateurs à « travailler pour consommer » (p. 21) ;
- la captation de productions bénévoles, où les consommateurs « coproduisent pour travailler » (p. 89) et consentent à céder gratuitement des productions pour le plaisir de travailler, de se sentir utiles, compétents ;
- la délégation du travail d'organisation, dans laquelle les consommateurs sont amenés à « trouver des solutions pratiques, socialement et subjectivement acceptables, à des contradictions qui apparaissent dans la coproduction » (p. 37).

Dans ce modèle, des situations de mise au travail sont ainsi distinguées selon le type d'activité et de consentement des consommateurs. Néanmoins, la question de la valeur y est absente. Elle est évoquée, à plusieurs reprises même, mais sans jamais être véritablement traitée.

C'est précisément sur ce point que le second modèle complète le premier. Portant plus spécifiquement sur « la création de valeur par le client » (Tiffon, 2009a), il précise de quelle valeur il s'agit, comment elle est créée, par qui, et en quoi les entreprises se l'accaparent. Au regard des questions qui nous intéressent, c'est donc à partir de ce modèle que nous examinerons la contribution des internautes aux *big data*.

2. La création de valeur par le client

La question de la valeur créée par le client sera ici appréhendée en deux temps. Le premier consiste à analyser les différentes manières dont le client contribue à l'efficacité productive ; le second à établir en quoi, et sous quelles conditions, ces contributions peuvent être considérées comme un travail, un travail productif, source de valeur pour les entreprises.

2.1. Utiliser le client pour accroître l'efficacité productive

S'appuyant sur une enquête de terrain réalisée dans des secteurs aussi divers que la grande distribution, la restauration rapide, les centres d'appel ou la kinésithérapie, ce modèle identifie quatre sortes de contribution des clients à l'efficacité productive.

2.1.1. La mise en astreinte

Avec la crise du fordisme (Durand, 1993 ; Boyer et Durand, 1993), les entreprises du secteur industriel se sont mises à « penser à l'envers » (Coriat, 1991) et à produire en flux tendu (Durand, 2004), c'est-à-dire à la demande, en suivant le principe du pilotage par l'aval, et en « juste-à-temps ». Ce retournement de la production a également eu lieu dans les services. Néanmoins, dans ces activités où il ne saurait y avoir de production sans coproduction et où, pour le dire autrement, la production dépend en partie de la présence et de la contribution des clients, un autre principe productif est également à l'œuvre, celui de l'externalisation des temps d'attente. En effet, l'intensification du travail des personnels de contact passant par une réduction des temps morts entre chaque client, les entreprises de service mettent tout en œuvre pour reporter ces temps d'attente sur le consommateur. Pour cela, elles ajustent toujours plus les effectifs aux périodes d'affluence, accroissent la polyvalence des salariés et, dans certains cas, notamment dans les petites structures, déparcellisent le travail pendant les périodes creuses de la journée (Tiffon, 2009b). Cette organisation du travail a pour corollaire une tendance à la systématisation des files d'attente. Ce temps d'attente, que supportent les clients, et qui participe directement de l'intensification du travail, c'est ce que j'appelle l'« astreinte client ».

2.1.2. La contribution managériale des clients

En sociologie du travail, les questions d'engagement dans le travail et de mobilisation des salariés ont fait couler beaucoup d'encre. Certains, dans le prolongement de l'anthropologie des techniques (Leroi-Gourhan, 1943, 1945), les appréhendent depuis l'activité (Bidet, 2011), le plaisir du faire ; d'autres depuis le prisme salarial, en insistant sur la peur du chômage (Coutrot, 1998), les contraintes émanant des nouvelles formes d'organisation du travail (Durand, 2004), l'évaluation individuelle et subjective des salariés (Dejours, 1998 ; Linhart, 1991, 1994), *via* le modèle de la compétence (Durand, 2004), l'espoir de promotions internes (Bernard, 2012a) ou l'idéologie managériale, qui mobilise le désir narcissique des salariés à des fins productives (Aubert et Gaulejac, 1991) et laisse entrevoir le travail comme un lieu de réalisation personnelle (Boltanski et Chiapello, 1999). Dans ces travaux, toutefois, les personnels de contact ne sont pas distingués des autres salariés, comme si la question du sens et de l'engagement dans le travail se posait de la même manière pour tous les travailleurs. Pourtant, les personnels de contact n'ont pas seulement affaire à des collègues et à une hiérarchie, mais aussi, et parfois surtout, à des clients. Or, ce que montre notre enquête, dans le prolongement de travaux sur la relation de service (notamment Jeantet, 2003 et Hanique, 2004), c'est que travailler, pour ces salariés, c'est justement être engagés dans une relation, un rapport social spécifique (Jeantet, 2003 ; Zarifian, 2013 ; Tiffon, 2013c), à la fois source de contraintes et de motifs d'engagement dans le travail (Tiffon, 2011b). En effet, d'un côté, par les marques de respect, de reconnaissance et de gratitude qu'ils adressent aux salariés, les clients participent grandement du sentiment d'être utile, de « rendre service ». En cela, ils suscitent l'engagement dans le travail des salariés. Mais, d'un autre côté, ils constituent également une contrainte extrêmement forte, en ce qu'ils contrôlent le travail qui s'opère sous leurs yeux, en particulier lorsqu'ils sont mis en situation d'attente (Tiffon, 2013d). La « contribution managériale des clients » renvoie précisément à ce double mécanisme, pour le moins ambivalent. Il est cependant certains cas, comme celui des caissières ou des équipiers du « McDo » – qui, en l'occurrence, sont particulièrement exposés à la pression du flux-client et ne jouissent pas de compétences spécialement reconnues – pour lesquels le contact avec le client est davantage source de contrôle que d'autonomie professionnelle, de même qu'il engendre davantage de marques d'ingratitude et de mépris (Tiffon, 2011c) que de reconnaissance et de valorisation sociale. La situation est quelque peu différente pour les « professionnels », comme les médecins ou les avocats, qui disposent non seulement de compétences reconnues et valorisées socialement, mais de dispositifs comme les salles d'attente pour exercer leur travail à l'abri des regards et d'autres formes de contrôle des clients (Tiffon, 2011b). Malgré des différences notables selon la place occupée dans l'espace social, les clients jouent donc un rôle non négligeable dans l'engagement dans le travail des salariés en *front office*.

2.1.3. L'intensification des interactions

En France, les activités de service constituent l'un des terrains privilégiés des analyses dites « interactionnistes ». Reprenant le cadre conceptuel élaboré par Erving Goffman dans *Asiles*

(1968), nombre de recherches s'évertuent à décrire les manières dont les clients et les personnels de contact co-construisent des solutions techniques, des accords marchands et des civilités dans le cours même de la relation de service. Dans ces travaux, toutefois, l'analyse des interactions est souvent déconnectée du rapport salarial. Comme si la co-construction d'accords prévalait sur les processus de rationalisation du travail. Pourtant, dans les services marchands, où la logique capitaliste de compression des coûts s'impose avec autant de force que dans l'industrie, les modes d'organisation du travail sont loin d'être sans incidence sur la nature des interactions. En tant que marchandises, les services n'échappent pas, en effet, à la recherche de réduction permanente du temps de travail socialement nécessaire à leur production. Qu'elle soit matérielle ou non, l'activité des salariés s'inscrit donc dans la même logique d'intensification du travail. À ceci près toutefois que, dans les services, rationaliser le travail des salariés en *front office*, chercher à optimiser le moindre déplacement, le moindre fait et geste n'a de sens que dans la mesure où, en face d'eux, les clients entrent, eux aussi, dans le flux de production. À travers le cas du « McDo », deux dispositifs permettant d'intensifier la contribution des clients ont pu être identifiés. Le premier est l'auto-contrôle client. Il renvoie au fait que, dans les files d'attente, les clients qui attendent exercent une certaine pression sur les clients indécis. « Gênés », ces derniers apprennent ainsi qu'il est tacitement convenu d'anticiper leur passage en caisse, en se renseignant sur l'offre et en choisissant leurs menus, avant même de passer commande. Ce faisant, le temps de passage en caisse est extrêmement court. Les échanges sont brefs et correspondent davantage à une « saisie de commande » qu'à un véritable diagnostic. Le second mécanisme renvoie, quant à lui, aux suggestions productives des personnels de contact, c'est-à-dire à tout ce qui, par leur attitude, exerce une pression sur les clients pour qu'ils commandent le plus rapidement possible. Au « McDo », les équipiers sont formés à cela : à la moindre hésitation des clients, ils doivent, quitte à leur couper la parole, formuler des suggestions, du type « menu *best of* ? » ou « frites-coca ? » de manière à orienter, mais surtout, à accélérer leur prise de commande. Lors du passage en caisse, les clients sont donc à la fois pressés, « poussés » par les clients qui attendent, et tirés par les salariés, qui, par leur attitude et leurs « suggestions productives », rendent manifestes les contraintes productives qui pèsent sur eux. Par ces deux mécanismes, les interactions s'intensifient. Et les clients en viennent à rationaliser leur manière de passer commande (Tiffon, 2007).

2.1.4. L'externalisation de tâches

La sociologie des services a étudié la coproduction en restant le plus souvent du côté des professionnels, réduisant ainsi l'analyse à ce que font les clients au cours de la relation de service. Or, il est une contribution des clients qui n'a pas nécessairement lieu en co-présence des salariés. Le plus souvent même, elle s'effectue seule. Cette contribution, c'est la prise en charge d'un travail qui était auparavant réalisé par des salariés. Elle fait suite à une externalisation de tâches et s'opère à travers deux dispositifs (Tiffon, 2013b).

Le premier, organisationnel et bien connu, est le *self-service*. Il renvoie au fait que les clients se servent directement en rayon et non plus par l'intermédiaire d'un vendeur. On le

retrouve non seulement dans la grande distribution alimentaire, mais aussi dans la restauration, avec les *fast-food* et les *self*, comme Flunch. On le retrouve également dans le prêt-à-porter, les magasins de bricolage ou les marchands de meubles, comme Ikea. Né dans la distribution alimentaire, il tend aujourd'hui à se généraliser à la plupart des activités de service. Sa spécificité, par rapport aux innovations organisationnelles observées dans l'industrie, est qu'il vise moins à intensifier le travail des salariés qu'à le réduire en substituant, sinon complètement, du moins en partie, les clients aux salariés en *front office*. Cela nécessite toutefois que les clients sachent comment faire. Pour cela, les entreprises mettent tout en œuvre pour les « socialiser » à leur organisation (Goudarzi, 2005). Les clients apprennent les nouvelles procédures de consommation et les intériorisent progressivement. Ils en viennent par exemple à débarrasser machinalement leur plateau, à se servir directement en rayon, à trouver seuls les produits qu'ils veulent, leur prix, voire leurs « avantages relatifs » par rapport aux autres produits. Avec le temps, ils finissent par savoir que chez Ikea, il faut, tout au long de leur parcours, relever les références des marchandises, noter les allées où elles se situent, pour ensuite aller eux-mêmes les chercher dans l'entrepôt avant leur passage en caisse. Bref, ils intériorisent peu à peu le fait de devoir d'abord essayer de se débrouiller seuls avant d'avoir éventuellement recours, en cas de problème, à l'aide de vendeurs.

Le second dispositif est technologique : il s'agit de ce que j'appelle l'« autonomation » – concept qui renvoie à la contraction des processus d'automatisation de la production et d'autonomisation de la contribution des clients. Si son principe générique est le même que celui du *self-service*, il se distingue de ce dernier dans la mesure où il renvoie à l'introduction de machines et est ainsi assimilé, de manière classique, au processus de substitution du capital au travail. Or, contrairement aux apparences, dans les services, les innovations technologiques ne se traduisent pas forcément par le remplacement de l'homme par des machines. Au contraire, dans certains cas, ces innovations constituent davantage un moyen de mettre au travail les clients, c'est-à-dire de remplacer du travail vivant salarié par une autre espèce de travail vivant, *a priori* gratuit. Ainsi en est-il des caisses automatiques, dans la grande distribution, au cinéma, dans les banques ou dans les stations-service. Ceux qui scannent les produits à la place des caissières, ce sont bien les clients, pas des machines (Bernard, 2012b ; Tiffon, 2011a ; Bernard et Tiffon, 2013). Ceux qui remplacent les salariés font le plein d'essence, renseignent le film à voir, indiquent la somme d'argent à retirer, etc. ; ce sont bien les clients, pas des machines. Ici, l'innovation tient donc moins à la substitution du capital au travail qu'à la création d'interfaces, suffisamment simples et didactiques, pour que les clients puissent, presque tâche pour tâche, se substituer aux salariés qui utilisaient jusqu'alors ces machines. Aussi Internet renvoie-t-il au même mécanisme : il permet de réduire non seulement la masse salariale, mais aussi les investissements fonciers, et autres frais, tout à fait considérables, que suppose le déploiement d'un parc d'agences sur tout un territoire. Les succès tout à fait notables dans le e-commerce (agences de voyages, commerces en ligne, etc.) tiennent essentiellement à cette compression des coûts de production. Qu'elle passe par l'introduction de machines ou non, la quatrième sorte de contribution des clients renvoie donc, de manière générique, à une externalisation de tâches, c'est-à-dire à la prise en charge par les clients d'une partie du travail auparavant réalisé par des salariés.

2.2. Contribuer à l'efficacité productive : un travail ?

Pour savoir en quoi ces contributions constituent un travail, il faut déjà s'entendre sur ce qu'est le travail. Or, compte tenu du caractère extrêmement polysémique de la notion, cela est loin d'être évident. Alain Cottereau (1994) n'en identifie pas moins de quatorze acceptions possibles que Michel Lallement (2001) regroupe en quatre catégories. Schématiquement, il y a :

- le *travail-souffrance* : travailler, c'est faire des efforts, se donner de la peine dans l'exercice d'une activité ;
- le *travail-praxis* : travailler, c'est y mettre du sien, dépasser l'incomplétude des prescriptions ;
- le *travail-statut* – ou *travail-intégration* : travailler, c'est s'inscrire dans une activité socialisée et socialisante ;
- le *travail-exploitation* : travailler, c'est créer de la valeur pour l'entreprise.

La première acception est beaucoup trop subjectiviste et relativiste. Dans cette perspective, est travail ce que les clients considèrent comme tel, c'est-à-dire tout et rien, et il suffit qu'ils y voient une dimension ludique ou récréative pour infirmer l'idée qu'ils travaillent – comme dans le cas des Datamatrix étudié par Aurélie Lachèze et Franck Cochoy (Cochoy et Lachèze, 2013). Quant aux deuxième et troisième acceptions, elles ne permettent pas, me semble-t-il, de véritablement distinguer le « travail » d'activités hors travail, comme les activités associatives ou de loisirs, qui elles aussi, peuvent être socialisées et socialisantes ou nécessiter que l'on y mette du sien. À mes yeux, donc, seule la dernière acception, qui associe le travail à la création de valeur, constitue un critère suffisamment robuste pour savoir si la contribution des clients peut être, ou non, considérée comme un travail. Mais dans cette perspective, la question devient : qu'est-ce que la valeur ? Et sous quelles conditions un travail crée-t-il de la valeur pour l'entreprise ? Questions pour lesquelles le cadre théorique marxiste m'est apparu particulièrement judicieux au regard des réponses qu'il apporte pour le cas des salariés.

2.3. Un travail productif, créateur de valeur pour les entreprises ?

S'inscrire dans une perspective marxiste nous positionne d'emblée dans une conception de la valeur, la conception marxienne de la valeur travail (Marx, 1867). Sujette à critiques, cette dernière a néanmoins le mérite, étant donné les questions qui nous intéressent ici, de placer le travail au fondement de la valeur, à la différence des marginalistes, et d'appréhender le profit, contrairement aux classiques, notamment Adam Smith (1776) et David Ricardo (1817), comme le résultat d'un procès d'extraction de valeur, d'une exploitation capitaliste des salariés. Je ne reviendrai pas ici sur la présentation de cette théorie de la valeur et des nombreux débats qu'elle a suscités ; de même que je ne reviendrai pas sur la théorie de l'exploitation, que chacun connaît et dont la présentation nécessiterait un détour trop important dans le cadre de ce chapitre. À ce stade de la démonstration, la question est plutôt de savoir en quoi les clients créent de la valeur pour les entreprises. En l'occurrence, la réponse n'est pas binaire et dépend du type de contributions.

Commençons par les deux premières : chez Marx, les managers occupent une fonction centrale, au sens où ils permettent d'optimiser la création de valeur de ceux qui produisent directement les marchandises, qu'il s'agisse d'un bien ou d'un service. Mais ils sont par nature improductifs. Leur fonction est juste, même si elle est essentielle, de tout mettre en œuvre pour accroître l'efficacité productive des (autres) salariés. Les deux premières sortes de contribution des clients – l'astreinte et la contribution managériale – ne sauraient donc être considérées comme un travail productif. Elles participent de l'efficacité productive des entreprises, permettent d'accroître leurs « performances », mais ne créent pas directement de la valeur. Les clients sont alors aussi improductifs que les managers.

Il en va autrement lorsqu'ils intensifient leur contribution et, surtout, prennent en charge une partie du travail auparavant réalisé par des salariés. Car, en ce cas, ils permettent de réduire le temps de travail socialement nécessaire à la production du service, donc sa valeur ; ce qui, dans la mesure où celle-ci devient inférieure à la valeur moyenne que l'on retrouve sur le marché, permet aux entreprises de tout à la fois baisser leurs prix et augmenter leur plus-value. Nous avons alors affaire à un mécanisme assez classique, celui de la « plus-value extra » (Marx, 1867, p. 236), qui a toutefois ceci de spécifique, ici, qu'il fait suite à la prise en charge par les clients d'une partie du travail auparavant réalisé par des salariés. D'où le concept que je propose de « néo-plus-value extra ».

Mais ce n'est pas tout. Cette plus-value, comme toute plus-value extra, ne dure qu'un temps. Si elle constitue une sorte de prime à l'innovation, avec le temps, les entreprises concurrentes sont contraintes, pour rester compétitives, d'adopter le même système de production. Autrement dit, si la concurrence sur le marché fonctionne – ce qui n'est pas toujours le cas ! – la plus-value extra est vouée à disparaître. En ce cas, toutefois, la baisse des prix donne alors lieu à une autre plus-value, une (néo-)plus-value relative. Si nous parlons de mise au travail, c'est donc que, par ces deux mécanismes – néo-plus-value extra et néo-plus-value relative –, les clients créent directement de la valeur que les entreprises s'accaparent.

3. Le cas des *big data*

Examinons à présent le cas des *big data*. Si l'on entend par là de grandes bases de données numériques qui stockent des informations sur les utilisateurs ou les clients à partir des traces qu'ils laissent d'eux sur Internet ; il en ressort que la première caractéristique de ce type de contribution est qu'elle se réalise sans co-présence. C'est là, me semble-t-il, un élément essentiel. Car cela signifie que les trois premières sortes de contribution qui viennent d'être exposées ne s'appliquent pas, de fait, au cas des *big data* :

- les temps d'attente des internautes ne permettent en aucun cas d'intensifier le travail des salariés ;
- leur contribution managériale est tellement indirecte et à ce point médiatisée qu'elle relève davantage de la rhétorique managériale sur le client et sa prétendue satisfaction que du rôle que ce dernier, fait de chair et d'os, joue réellement dans le rapport au travail des salariés ;
- le fait que les internautes aillent vite, ou non, n'a aucune espèce d'incidence sur le temps de travail des salariés et leur productivité.

La seule sorte de contribution qui pourrait alors se rapprocher du cas des *big data*, c'est donc la quatrième, celle de l'accroissement de la contribution des clients. Seulement, pour cela, encore faut-il que les internautes effectuent un « travail », disons une activité, auparavant réalisée par des salariés.

Or, qu'y avait-il avant ces *big data* ? Comment les entreprises avaient-elles accès à l'avis et aux attentes des consommateurs ? Ces informations étaient le plus souvent collectées à travers des questionnaires de satisfaction – comme continuent d'ailleurs de le faire un certain nombre d'entreprises. Ces informations étaient donc déjà coproduites. Les clients répondaient à des questions, exprimaient leurs opinions, donnaient accès à leurs goûts, à leurs préférences. Mais des salariés concevaient ces questionnaires, en analysaient les réponses et en tiraient des enseignements, en termes de conception de l'offre et d'ajustement de la communication.

Qu'y a-t-il de différent aujourd'hui ? Avec les *big data*, qui fait quoi ? Sont-ce les internautes qui conçoivent les dispositifs de traçabilité sur Internet ? Sont-ce les internautes qui en analysent les données ? Sont-ce ces mêmes internautes qui en tirent des enseignements pour renouveler l'offre et ajuster la communication des entreprises ? *A priori*, dans la plupart des cas, leur contribution se situe à un autre niveau. Ce qu'ils font, c'est ce qu'ils ont toujours fait, à savoir donner accès à leurs pratiques et à leurs attentes. Ce qui a changé, c'est juste leur manière d'y donner accès. Avant, c'était en répondant à des questionnaires. Aujourd'hui, c'est de plus en plus, et sans nécessairement en avoir conscience, en laissant des traces d'eux, de ce qu'ils font, de ce qu'ils aiment, sur Internet, en navigant sur Google, en postant des photos, des musiques ou des commentaires sur Facebook.

En somme, dans la mesure où l'activité des internautes ne se substitue pas à celle de salariés, il serait impropre, au regard du modèle exposé, de parler de mise au travail pour décrire leur contribution à la constitution des grandes bases de données numériques dont disposent les entreprises. S'il est des cas où les clients-utilisateurs sont mis au travail, *via* Internet, il en va donc autrement pour ce cas précis : effectivement, les internautes sont au cœur du *business model* de certaines entreprises, comme Google ou Facebook, qui « ne pourraient pas exister sans ces utilisateurs » qui laissent des traces d'eux « sans compensations financières » (Fuchs, 2013, p. 213) ; effectivement, la « fiscalité du numérique » est une question cruciale à laquelle il semble essentiel de trouver des réponses pour que les entreprises de ce secteur ne puissent plus échapper aux fiscalités nationales des pays « utilisateurs » (Collin et Colin, 2013) ; mais ce n'est pas pour autant que la contribution des internautes correspond à un « travail gratuit » – comme le soutiennent Pierre Collin et Nicolas Colin, pour justifier leur proposition d'imposer les firmes de l'économie numérique selon la territorialité des utilisateurs (Collin et Colin, 2013) – ou à un travail directement producteur de valeur, au sens marxiste du terme (Fuchs, 2013 et 2014).

S'il fallait qualifier ces *big data*, je dirais plutôt qu'elles correspondent surtout à un processus d'automatisation d'une partie du travail auparavant réalisé par des salariés. Même si, bien sûr, l'ampleur de ces bases de données comme les potentialités qu'offre aujourd'hui Internet dépassent de très loin le travail des équipes marketing ayant réa-

lisé les premières études de marché ; même si, évidemment, l'activité en tant que telle n'a plus rien à voir ; il me semble que, d'un point de vue fonctionnaliste, ces *big data* correspondent en effet à une forme d'automatisation d'une des fonctions du marketing et donc, d'une partie du travail qui, de façon générique, pouvait être, et est encore parfois, réalisé au sein de ces services.

Conclusion

Au regard de ce qui vient d'être dit, l'enjeu des *big data* se pose moins, me semble-t-il, en termes de mise au travail des clients que de spoliation de données personnelles, ensuite marchandisées, et donc, de contrôle, de surveillance et, par conséquent, de débat démocratique à engager sur la protection de ces données et les limites à donner à cette nouvelle espèce de « mise sur écoute » généralisée sur la Toile. Autrement dit, j'aurais plutôt tendance à penser que, de ce point de vue, le cas des *big data* appelle davantage à une analyse foucaldienne que marxiste. Car sur Internet, on le sait, chaque fait et geste laisse une trace qui alimente des bases de données, à partir desquelles les entreprises, *via* des algorithmes dont la puissance de calcul évolue de façon exponentielle, ajustent toujours plus leur offre et leur communication.

Si, du point de vue des entreprises, cette connaissance de plus en plus fine des pratiques, des représentations et des attentes des consommateurs est déterminante pour assurer des débouchés à leur production et se développer ; en revanche, pour les consommateurs, cette masse de données dont disposent les entreprises ne fait que renforcer leur « vulnérabilité », c'est-à-dire leur incapacité à se déprendre de l'emprise du marché, des frustrations qu'il suscite et du désir d'achats, plus ou moins compulsif, qu'il alimente.

Il y a donc là, me semble-t-il, quelque chose qui, dans les *big data*, entendues comme mécanisme d'extraction de données en vue d'une utilisation marketing, relève de l'aliénation, par la consommation, et donc, aussi, par le travail ; car le fameux cercle vertueux fordien de référence, celui de la productivité, de la croissance, de l'élévation des niveaux de vie et de la consommation de masse, est aussi celui, par certains aspects, vicieux, de l'assujettissement de l'homme aux finalités du mode de production capitaliste, celui dans lequel la centralité du travail devient aussi indiscutée et indiscutable qu'elle nourrit l'ivresse de la consommation. Si l'approche marxiste peut être utile à l'analyse des *big data*, c'est donc moins, me semble-t-il, en termes d'exploitation que d'aliénation. Mais c'est là une hypothèse déductive qui, dans une approche sociologique, mériterait investigation.

Références

Aubert N. et Gaulejac V. de (1991), *Le Coût de l'excellence*, Paris, Seuil.

Bendapudi N. et Leone R.P. (2003), « Psychological implications of customer participation in coproduction », *Journal of Marketing*, n° 67, p. 14-28.

Bernard S. (2012a), « La promotion interne dans la grande distribution : la fin d'un mythe ? », *Revue française de sociologie*, n° 2, p. 259-291.

Bernard S. (2012b), *Travail et automatisation des services. La fin des caissières ?*, Toulouse, Octarès.

Bernard S. et Tiffon G. (2013), « De l'automatisation des caisses à la recomposition du travail des caissières », *in* Durand J.-P., Moatty F. et Tiffon G. (dir.), *L'Innovation dans le travail*, Toulouse, Octarès, p. 77-89.

Bernard S., Dujarier M.-A. et Tiffon G. (dir.) (2011a), *L'activité des clients : un travail ?*, *Sciences de la société*, n° 82, http://sds.revues.org/2012.

Bernard S., Dujarier M.-A. et Tiffon G. (2011b), « L'hypothèse de la mise au travail des clients », *Sciences de la société*, n° 82 (*L'activité des clients : un travail ?*), p. 3-19, http://sds.revues.org/2016 (mis en ligne le 7 septembre 2015, consulté le 27 février 2017).

Bidet A. (2011), *L'Engagement dans le travail. Qu'est-ce que le vrai boulot ?*, Paris, PUF.

Boltanski L. et Chiapello E. (1999), *Le Nouvel Esprit du capitalisme*, Paris, Gallimard.

Bowen D.E. (1986), « Management customers as human resources in service organizations », *Human Resource Management*, vol. 25, n° 3, p. 371-383.

Boyer R. et Durand J.-P. (1993), *L'Après-Fordisme*, Paris, Syros.

Chase R.B. et Dasu S. (2001), « Want to perfect your company's service? Use behavioural science », *Harvard Business Review*, vol. 79, n° 6, p. 78-84.

Cochoy F. et Lachèze A. (2013), « L'usage des codes-barres 2D comme *self-marketing* : travail du consommateur ou curiosité en jeu ? », *Sciences de la société*, n° 82, p. 59-79, http://sds.revues.org/2038 (mis en ligne le 7 septembre 2015, consulté le 21 février 2017).

Collin P. et Colin N. (2013), « Mission d'expertise sur la fiscalité de l'économie numérique », rapport remis en janvier 2013 au ministre de l'Économie et des Finances, au ministre du Redressement productif, au ministre délégué chargé du Budget et à la ministre déléguée chargée des Petites et Moyennes Entreprises, de l'Innovation et de l'Économie numérique.

Coriat B. (1991), *Penser à l'envers*, Paris, C. Bourgeois.

Cottereau A. (1994), « Théories de l'action et notion de travail », *Sociologie du travail*, n° 36, p. 73-89.

Coutrot T. (1998), *L'Entreprise néo-libérale. Nouvelle utopie capitaliste ?*, Paris, La Découverte.

Dejours C. (1998), *Souffrance en France*, Paris, Seuil.

Dujarier M.-A. (2008), *Le Travail du consommateur*, Paris, La Découverte.

Durand J.-P. (2004), *La Chaîne invisible. Travailler aujourd'hui : flux tendu et servitude volontaire*, Paris, Seuil.

Durand J.-P. (dir.) (1993), *Vers un nouveau modèle productif ?*, Paris, Syros.

Eiglier P. (2004), *Marketing et stratégie des services*, Paris, Economica.

Eiglier P. et Langeard E. (1987), *Servuction. Le marketing des services*, Paris, Mc Graw-Hill.

Fuchs C. (2013), « Class and exploitation on the Internet », *in* Scholz T. (dir.), *Internet as Playground and Factory*, New York, Routledge, p. 211-224.

Fuchs C. (2014), *Digital Labour and Karl Marx*, New York, Routledge.

Goffman E. (1968), *Asiles. Étude sur la condition sociale des malades mentaux*, Paris, Éditions de Minuit.

Goudarzi K. (2005), *Le Concept de socialisation organisationnelle des clients dans les entreprises de service, thèse de doctorat en marketing*, Université d'Aix-Marseille III.

Hanique F. (2004), *Le Sens du travail. Chronique de la modernisation au guichet*, Paris, Érès.

Jeantet A. (2003), « "À votre service !" La relation de service comme rapport social », *Sociologie du travail*, nº 45, p. 191-209.

Kelley S. W., Donnely J.H. et Skinner S. J. (2001), « Customer participation in service production and delivery », *Journal of Retailing*, p. 315-335.

Lallement M. (2001), « *Daedalus laborans* », *Revue du MAUSS*, nº 18, p. 29-49.

Lapert D. (2005), *Le Marketing des services*, Paris, Dunod.

Larsson R. et Bowen D.E. (1989), « Organization and customer: managing design and coordination of services », *Academy of Management Review*, vol. 14, nº 2, p. 213-233.

Leroi-Gourhan A. (1943), *Évolution et techniques*, vol. 1 : *L'Homme et la Matière*, Paris, Albin Michel, 1971.

Leroi-Gourhan A. (1945), *Évolution et techniques*, vol. 2 : *Milieu et techniques*, Paris, Albin Michel, 1973.

Linhart D. (1991), *Le Torticolis de l'autruche. L'éternelle modernisation des entreprises françaises*, Paris, Seuil.

Linhart D. (1994), *La Modernisation des entreprises*, Paris, La Découverte.

Lovelock C.H. et Young R.F. (1979), « Look to customers to increase productivity », *Harvard Business Review*, nº 57, p. 168-178.

Marx K. (1867), *Le Capital. Livre I*, Paris, Flammarion, coll. « Champs », 1985.

Mills P.K. et Morris J.H. (1986), « Client as partial employees of service organizations: role development in client participation », *The Academy of Management Review*, vol. 11, nº 4, p. 726-735.

Mills P. K., Chase R. B. et Margulies N. (1983), « Motivating the client/employee system as a service production strategy », *The Academy of Management Review*, vol. 8, nº 2, p. 301-310.

Prahalad C.K. et Ramaswamy K. (2000), « Co-opting customer competence », *Harvard Business Review*, p. 79-87.

Ricardo D. (1817), *Principes de l'économie politique et de l'impôt*, Paris, Flammarion, coll. « Champs », 1992.

Smith A. (1776), *Recherches sur la nature et les causes de la richesse des nations*, Paris, Economica, 2005.

Tiffon G. (2007), « La microsociologie n'est pas de l'individualisme », *in* Durand J.-P. et Gasparini W. (dir.), *Le Travail à l'épreuve des paradigmes sociologiques*, Toulouse, Octarès, p. 85-97.

Tiffon G. (2009a), *La Création de valeur par le client*, thèse de doctorat en sociologie, Université d'Évry.

Tiffon G. (2009b), « Le flux pressé tendu : de la rationalisation de l'organisation du travail à la généralisation des files d'attente », *in* Appay B. et Jefferys S. (dir.), *Restructurations, précarisation, valeurs*, Toulouse, Octarès, p. 157-168.

Tiffon G. (2011a), « Ces automates qui nous font travailler », *La Pensée*, vol. 366, p. 63-76.

Tiffon G. (2011b), « La contrainte client. Une analyse comparée des caissières et des kinésithérapeutes », *SociologieS*, rubrique « Premiers textes », http://sociologies.revues.org/3444 (mis en ligne le 11 avril 2011, consulté le 27 février 2017).

Tiffon G. (2011c), « Quand le comportement des clients fait violence : le cas des caissières de la grande distribution spécialisée », *in* Dressen M. et Durand J.-P. (dir.), *La Violence au travail*, Toulouse, Octarès.

Tiffon G. (2013a), *La Mise au travail des clients*, Paris, Economica, coll. « Études sociologiques ».

Tiffon G. (2013b), « La contre-externalisation comme innovation de procédé de service. Un ressort de la rationalisation du travail propre aux activités de service », *in* Goussard L. et Sibaud L. (dir.), *La Rationalisation dans tous ses états*, Paris, L'Harmattan.

Tiffon G. (dir.) (2013c), « Relation de service, rapport social de service : quelle grille d'analyse ? », *La Nouvelle Revue du travail*, n° 2, http://nrt.revues.org/759 (mis en ligne le 17 mars 2013, consulté le 27 février 2017).

Tiffon G. (2013d), « La pression du flux client », *in* Bercot R. et Rahou N. (dir.), *Le Travail de service*, Paris, Éd. de l'Agence nationale pour l'amélioration des conditions de travail (ANACT).

Zarifian P. (2013), « Rapport social de service, client et valeur », *La Nouvelle Revue du travail*, n° 2, http://nrt.revues.org/737 (mis en ligne le 30 mars 2013, consulté le 27 février 2017).

II. *Big data* et configurations sociales en mouvement

La « science des données » à la conquête des mondes sociaux : ce que le « Big Data » doit aux épistémologies locales

Éric Dagiral
Maître de conférences en sociologie à l'université Paris Descartes
et chercheur au CERLIS (Centre de recherche sur les liens sociaux)

Sylvain Parasie
Maître de conférences en sociologie à l'université Paris-Est / Marne-la-Vallée
et chercheur au LISIS (Laboratoire interdisciplinaire sciences, innovations, sociétés)

Au milieu des années 1950, le pionnier de l'informatique Howard Aiken était loin de penser que l'ordinateur trouverait une place en dehors des centres de recherche[1]. « S'il s'avérait, écrivait-il, que la logique de base d'une machine conçue pour la résolution numérique des équations différentielles coïncidât avec la logique d'une machine conçue pour établir les factures d'un grand magasin, je considérerais ce fait comme la coïncidence la plus extraordinaire que j'aie jamais vue[2] » (Ceruzzi, 1993). Soixante ans plus tard, non seulement l'informatique est utilisée dans la plupart des mondes sociaux, mais ce sont aussi désormais de plus en plus les façons de calculer issues des mondes informatiques et statistiques qui pénètrent les mondes sociaux les plus divers.

Qu'il s'agisse du marketing, de la santé, des médias, du sport ou même de l'industrie automobile, les mondes sociaux les plus variés en viennent à mobiliser aujourd'hui des algorithmes et des modèles statistiques qui leur sont en apparence éloignés. Ces dernières années, le phénomène du « Big Data » a à la fois rendu visible et contribué à étendre la mobilisation, par les mondes sociaux les plus divers, de technologies informatiques et statistiques pour traiter d'immenses volumes de données souvent hétérogènes. La remarque de Howard Aiken reste pourtant d'actualité. Comment expliquer que des technologies de calcul élaborées par des statisticiens et des informaticiens en viennent à « coïncider » avec les préoccupations de médecins, de *marketers*, d'industriels, de journalistes et même d'entraîneurs sportifs ? Sur quels fondements pratiques et cognitifs une telle conquête s'établit-elle ?

Face à une telle question, les promoteurs du Big Data soulignent d'abord que la plupart des organisations et des individus sont aujourd'hui confrontés à un ensemble de problèmes pratiques communs liés à la disponibilité de vastes ensembles de données souvent « sales », peu, voire non structurées, et provenant de sources hétérogènes. Mais plus profondément, expliquent-ils, les technologies actuelles permettraient surtout une rupture de nature épistémologique. Alors que chaque monde social est marqué par un ensemble

1. Nous remercions les coordinateurs de cet ouvrage ainsi que Dominique Cardon et Jean-Philippe Cointet pour leurs précieux conseils.
2. Sauf mention contraire, les citations sont traduites en français par les auteurs du chapitre [NdE].

de préjugés ou d'hypothèses jamais interrogées, les techniques du Big Data offriraient un retour à une forme de connaissance plus inductive (Anderson, 2008). Comme l'écrivent Mayer-Schönberger et Cukier,

> […] si le Big Data est susceptible d'offrir un nouveau regard et de nouvelles connaissances, c'est précisément parce qu'il n'est pas gêné par les conceptions traditionnelles ou les préjugés qui se cachent derrière les théories d'un domaine spécifique (Mayer-Schönberger et Cukier, 2013, p. 71).

L'attrait pour les perspectives analytiques de type *big data* résiderait donc surtout dans la forme de connaissance plus inductive que celles-ci porteraient – laquelle rendrait possible la conception de produits ou de services valorisables économiquement.

Les chercheurs en sciences sociales ont accueilli ces arguments avec un intérêt teinté d'un certain scepticisme. Le Big Data, ont-ils avancé, doit être étudié comme un phénomène technologique et culturel (Boyd et Crawford, 2012), mais cela fait bien longtemps que l'on fait face à des données massives et hétérogènes (Strasser, 2012 ; Grier, 2005). En outre, ils ont remis en cause l'argument selon lequel le Big Data marquerait une rupture épistémologique. Pointant la résurgence de discours empruntés d'une approche inductive naïve, ils ont rappelé que les données sont toujours des constructions à la fois sociales et politiques (Gitelman et *al.*, 2013), et que des traditions scientifiques bien établies visaient déjà à accumuler un grand nombre d'éléments empiriques sans formuler d'hypothèses de départ (Strasser, 2012). Quoique pertinentes, ces critiques échouent pourtant à rendre compte de l'extension du phénomène. La question reste donc entière : la mobilisation de ces technologies de traitement de données dans une diversité de mondes sociaux s'explique-t-elle par la séduction d'un modèle de connaissance plus inductif ? L'extension des *big data* correspond-elle à une remise en cause des épistémologies spécifiques à chaque monde social et à ses segments professionnels (Strauss, 1992) ?

Considéré depuis la sociologie des sciences et des techniques, l'argument selon lequel les technologies des *big data* seraient porteuses d'une remise en cause des épistémologies locales est plutôt contre-intuitif – si on entend par « épistémologies locales » des façons de connaître imbriquées dans des mondes sociaux (Knorr-Cetina, 1999). Plusieurs théories ont mis l'accent sur la plasticité des objets technologiques et scientifiques, laquelle rendrait possible une pluralité d'interprétations à l'intérieur des mondes sociaux qui s'en emparent (Star et Griesemer, 1989). Analysant la mise au point de la bombe nucléaire dans l'Amérique des années 1940, l'historien Peter Galison a ainsi montré de quelle manière les méthodes statistiques de Monte Carlo ont été au cœur de la collaboration entre des mondes scientifiques et industriels qui les interprétaient chacun d'une façon particulière (Galison, 1996).

Pour être en mesure de trancher cette question, il est impératif, croyons-nous, de déplacer notre regard. Mettons de côté le terme *big data*, auquel les professionnels n'ont souvent recours que pour chercher à convaincre des clients ou sensibiliser le public aux enjeux économiques, industriels et politiques du traitement de données. Concentrons-nous plutôt sur une expression qui est à la fois davantage utilisée par les praticiens et qui renvoie à des savoirs et à des technologies beaucoup plus situés : la « science des données » ou *data science*. Apparue au tout début des années 2000, cette expression désigne un ensemble de pratiques, de savoirs et de technologies situées au croisement des mondes de l'informatique et des statistiques.

Elle s'applique aussi plus récemment à un métier – celui de *data scientist* – aujourd'hui reconnu par plusieurs institutions du marché du travail aux États-Unis comme en Europe[3]. Embauché par les entreprises du Web et les organisations les plus diverses, ce praticien est chargé de concevoir des *data products*, c'est-à-dire des services ou des produits élaborées à partir d'importants volumes de données souvent peu structurées.

Une fois ce déplacement opéré, nous verrons qu'il est plus facile de décrire et d'expliquer l'extension de ces savoirs et technologies à des mondes sociaux variés. Là où, bien souvent, les chercheurs critiquent la prétention des *big data* à s'appliquer à des mondes sociaux très différents les uns des autres, nous verrons au contraire que leur capacité de diffusion procède bien plutôt de leur relative ouverture aux épistémologies propres aux différents mondes sociaux concernés.

L'enquête s'appuie ici sur trois types de matériaux. En premier lieu, nous avons analysé la littérature académique dans le domaine des statistiques et de l'informatique depuis le début des années 1960. Nous avons ainsi pu reconstituer l'émergence progressive d'un domaine associé à la « science des données ». En deuxième lieu, nous avons étudié une partie des éléments logiciels constitués dans le cadre du projet *open source* R – qui constitue une référence majeure pour les praticiens des *data sciences*. Enfin, nous avons conduit une enquête par entretiens auprès d'une dizaine de *data scientist*s nord-américains engagés dans les mondes du journalisme d'une part, du corps, de la santé et du bien-être d'autre part.

Notre propos s'organise de la façon suivante. Dans une première partie, nous montrons que la « science des données » résulte de la revalorisation d'une tradition longtemps minoritaire dans le monde des statistiques. Cette revalorisation découle d'une fragilisation professionnelle des statisticiens face à l'essor des pratiques d'analyse de données rendues possibles par les développements de l'informatique depuis les années 1960. Dans une deuxième partie, nous analysons les principaux traits de la science des données depuis le début des années 2000 et leur diffusion au sein d'une pluralité de mondes sociaux. Enfin, nous consacrons la troisième partie à l'analyse de la réception de cette « science des données » et de sa mise à l'épreuve conjointe dans deux mondes particuliers – le monde du journalisme et celui du corps, de la santé et du bien-être.

1. Petite histoire de la « science des données »

Ce qu'on appelle aujourd'hui la science des données désigne un certain assemblage de technologies et de savoirs situés au croisement des statistiques et de l'informatique. Cet assemblage trouve son origine dans l'Amérique des années 1960, au moment où le traitement de données connaît un essor dans les entreprises et les universités. Au sein du monde de la statistique universitaire américaine apparaît alors un mouvement d'ouverture des statistiques à l'informatique et aux préoccupations des mondes sociaux.

3. En 2013, l'APEC (Association pour l'emploi des cadres) a fait figurer le poste de « *data scientist* » ou « scientifique de la donnée » dans la liste des « métiers en émergence ». Cf. APEC, *Les métiers en émergence*, hors-série, « Les référentiels des métiers cadres », mai 2013, p. 65.

1.1. Un mouvement de statisticiens dans l'Amérique des années 1960

John Tukey est la figure majeure d'un mouvement qui prend naissance au sein de la statistique américaine du début des années 1960. Ce mathématicien de formation officie à la fois à l'université de Princeton et dans les laboratoires de recherche de la compagnie AT&T – les Bell Labs –, où il développe des méthodes statistiques et promeut la conception d'outils statistiques informatisés. En 1962, il publie un long article programmatique intitulé « Le futur de l'analyse de données », dans lequel il explique que les statistiques doivent s'ouvrir bien au-delà de leur dimension strictement mathématique (Tukey, 1962).

Son argumentation prend la forme suivante. Il enjoint d'abord à ses collègues statisticiens de s'intéresser à des problèmes pratiques émanant de professionnels d'horizons divers. Lui-même a participé à des projets d'analyse de données en prise directe avec des mondes précis : lors des élections présidentielles de 1960, il dirige une équipe qui prédit, pour le compte de la chaîne NBC, le résultat du vote à partir des premiers dépouillements (McGrayne, 2011, p. 163-175). Il incite ensuite les statisticiens à analyser des données produites par les professionnels des mondes concernés, qui sont donc souvent très incomplètes – comme dans le cas des données électorales. À ses yeux, cela implique que les statisticiens doivent s'intéresser de près à l'ordinateur, et non plus seulement aux mathématiques. Ils y trouveront, explique-t-il, à la fois de nouveaux outils d'analyse et de nouvelles façons de représenter visuellement des données.

Mais Tukey met en avant un argument plus fondamental encore. En se convertissant à l'« analyse de données », les statisticiens modifieront leur propre rapport à la connaissance. Plutôt que de chercher à appliquer des modèles mathématiques en vue de la plus grande précision possible, le statisticien devra davantage s'interroger sur les « bonnes questions », quitte à en rabattre sur la précision mathématique :

> La maxime la plus importante à respecter pour l'analyse de données – et que de nombreux statisticiens ont dédaignée – est celle-ci : « Mieux vaut de loin une réponse approximative à la bonne question, laquelle est souvent vague, plutôt qu'une réponse exacte à la mauvaise question, laquelle peut toujours être précise. » (Tukey, 1962)

Quelques années plus tard, Tukey fait la distinction entre deux paradigmes distincts de l'analyse de données (Tukey, 1977). D'une part, l'« analyse confirmatoire des données » (*confirmatory data analysis*) consiste, à partir d'une question bien définie, à mettre en place un protocole d'enquête et à collecter des données que l'on analyse ensuite pour en tirer une conclusion. À ses yeux, ces opérations sur les données peuvent être informatisées, c'est-à-dire en grande partie « routinisées ». D'autre part, l'« analyse exploratoire des données » (*exploratory data analysis*) vise plutôt à laisser ouverte la question sur laquelle portera la recherche et à manipuler un ensemble d'outils de représentation des données pour y déceler progressivement la question pertinente et les types d'analyses possibles. Si ces deux paradigmes sont à ses yeux complémentaires, il est urgent de développer l'analyse exploratoire des données :

> Pour suggérer rapidement en quoi consiste l'analyse exploratoire des données, je dirais qu'il s'agit (1) d'une attitude, (2) d'une flexibilité et (3) de papiers millimétrés (ou de transparents, ou des deux à la fois). Aucun catalogue de techniques ne peut transmettre une disposition à chercher ce qui peut être observé, que cela ait été anticipé ou non. C'est pourtant là qu'est le cœur de l'analyse exploratoire de données. (Tukey, 1980, p. 24)

Dès les années 1960, un mouvement apparaît alors au sein du monde des statistiques américaines en faveur de l'analyse de données[4]. Celui-ci défend l'importance de l'exploration aux différentes étapes d'une analyse faite d'essais, d'allers et retours, recourant aux visualisations pour chercher à identifier des relations entre des données, puis à s'interroger sur les significations de ces relations et, éventuellement, à recommencer l'analyse à partir de nouveaux éclairages. Ce mouvement correspond également à l'ouverture des statisticiens à des considérations issues du monde des entreprises et d'une diversité de mondes sociaux.

1.2. L'informatique, vecteur de diffusion de l'analyse de données

En pratique, c'est la diffusion de logiciels statistiques qui étend le traitement de données en dehors du monde de la statistique universitaire. À partir de la fin des années 1960 apparaît en effet une offre commerciale qui s'adresse à des personnes qui ne maîtrisent pas nécessairement le contenu mathématique des modèles statistiques. Il s'agit d'abord de chercheurs non statisticiens, puis de membres des administrations et des entreprises. À tel point que le terme même de *statisticien* se met à désigner des personnes qui exercent des activités dans des domaines professionnels très différents (comptabilité, assurance, banque, etc.).

Conçus pour des ordinateurs *mainframe* (ordinateurs de grande puissance de calcul), les trois premiers logiciels s'adressent au départ à des usagers différents. Le premier s'appelle BMDP (BioMeDical Package) et il est développé à l'université de Los Angeles au sein d'un centre de recherche biomédicale. Il est destiné aux chercheurs travaillant sur la santé. Le deuxième logiciel statistique cible les chercheurs en sciences sociales : il se nomme SPSS (Statistical Package for the Social Sciences) et apparaît en 1968 à l'université de Chicago. Appelé SAS (Statistical Analysis System), le troisième logiciel est produit la même année à l'université de Caroline du Nord, et s'adresse spécifiquement aux entreprises. Le statisticien Jan De Leeuw explique que cette première génération de logiciels reposait sur un répertoire commun de techniques statistiques :

> Les spécialistes de la recherche en santé, les chercheurs en sciences sociales et les clients dans les entreprises, tout le monde avait besoin d'un répertoire standard de techniques statistiques, auquel s'ajoutaient des méthodes spécialisées qui étaient importantes dans chaque domaine. Donc, les paquets logiciels se différenciaient, même si les éléments de base étaient *grosso modo* les mêmes. (De Leeuw, 2011, p. 9)

Au cours des années 1980, le traitement de données connaît un essor important dans une diversité de mondes sociaux : d'une part, parce que les firmes qui commercialisent ces logiciels statistiques élargissent leur clientèle en tirant partie de la diffusion de l'ordinateur personnel ; d'autre part, parce que l'informatisation des grands organismes au cours de la

4. Dès 1970, l'économiste français Edmond Malinvaud repère ce « mouvement pour l'analyse de données » qu'il interprète comme « visant surtout à un élargissement de la méthodologie statistique, un élargissement qui s'avère opportun même s'il suppose que soient réduites, dans un premier temps, les ambitions de rigueur auxquelles la statistique classique a pu nous habituer » (Malinvaud, 1970, p. 7).

décennie s'accompagne d'une intensification de la production de données numériques, susceptibles de faire l'objet à leur tour d'analyses menées à l'aide de tels logiciels.

L'analyse de données réalisée au sein des organismes au moyen de ces logiciels prend alors des formes différentes[5]. Les modèles de régression connaissent ainsi une extension importante. Mais certains logiciels offrent aussi des fonctionnalités graphiques qui rendent possible une certaine exploration des données que John Tukey avait appelée de ses vœux vingt ans plus tôt. Un autre rapport aux données, moins directement centré sur l'identification de facteurs explicatifs, devient également possible.

En marge des firmes qui commercialisent SPSS et SAS auprès d'un nombre croissant d'organisations, des logiciels plus confidentiels apparaissent aux États-Unis à partir de la fin des années 1980. Conçus par des chercheurs des Bell Labs, ceux-ci visent à permettre l'analyse exploratoire de données imaginée par John Tukey. C'est notamment le cas du langage de programmation statistique appelé « S », qui est alors seulement utilisé dans une partie du monde académique américain. À la fin des années 1990, ce langage devient « R » et connaît un succès très important (De Leeuw, 2011). Il s'impose dans le monde académique nord-américain, avant de s'étendre à l'enseignement des statistiques et à bien d'autres mondes sociaux[6]. Une enquête réalisée en 2011 par une société privée auprès d'un millier d'analystes indiquait que R était le logiciel le plus utilisé dans le monde académique, les entreprises, l'administration et les ONG[7].

Depuis les années 1960, la diffusion de l'informatique et des logiciels statistiques a ainsi participé à l'avènement progressif de « l'analyse de données » telle que la défendait John Tukey. Une variété toujours plus grande de mondes sociaux s'est mise à mobiliser les statistiques pour traiter, et même explorer, des données souvent hétérogènes.

1.3. La « science des données », une solution à la crise des statistiques

L'expression « science des données » apparaît au tout début des années 2000 comme une manière, pour les statisticiens universitaires, de combattre ce qu'ils identifient alors comme une crise de leur domaine. Nombreux sont en effet ceux qui dressent un constat sévère de la situation des statistiques au tournant des années 2000 (Chambers, 1999). D'abord, la profession est désormais concurrencée par de nouvelles spécialités scientifiques :

> Nous ne sommes plus les seuls sur le marché. Jusqu'à récemment, pour quelqu'un qui était intéressé par l'analyse de données, les statistiques étaient l'un des très rares domaines adéquats dans lequel il était possible de travailler. Ce n'est plus le cas. Il y a aujourd'hui plusieurs autres sciences passionnantes qui sont orientées vers les données, et qui sont

5. Alors que l'usage des statistiques dans les institutions publiques a fait l'objet de travaux historiques majeurs, l'histoire des pratiques statistiques en entreprise reste très peu documentée.

6. À tel point que les logiciels SAS et SPSS, qui étaient dominants dans les entreprises, ont été contraints d'inclure des interfaces avec R (De Leeuw, 2011).

7. Cf. D. Smith (2013), « R users: Be counted in Rexer's 2013 Data Miner Survey », *Revolution Analytics Blog*, 30 janvier 2013. http://blog.revolutionanalytics.com/2013/01/r-users-be-counted-in-rexers-2013-data-miner-survey.html (consulté le 13 novembre 2014).

> en concurrence avec nous pour attirer des clients, des étudiants, des boulots et même nos propres statisticiens. (Friedman, 2001)

Quelles sont ces « autres sciences passionnantes » qui viennent ainsi priver les statisticiens professionnels de leurs forces vives ? Ceux-ci ont alors surtout à l'esprit le *data mining*, un domaine interdisciplinaire issu de l'informatique qui se consacre à l'analyse de vastes corpus de données. S'il est alors jugé particulièrement dangereux pour les statistiques, c'est que ce domaine vise d'emblée un grand nombre d'entreprises, en tirant profit des volumes considérables de données que celles-ci collectent auprès de leurs clients. Or les statistiques ne jouent qu'un rôle marginal dans ce domaine en plein essor :

> La communauté du *data mining* devra peut-être modérer son amour du « *big* » : les méthodes d'échantillonnage, qui ont une longue tradition en statistiques, peuvent être utilisées avec profit de façon à améliorer la précision et à réduire les besoins en capacité de calcul. (Friedman, 1997)

Plus généralement, ces statisticiens s'inquiètent de voir toutes les activités associées à l'analyse de données être directement contrôlées par les informaticiens et les firmes informatiques. Collecter les données, gérer les bases de données, traiter les données et les visualiser : toutes ces opérations sont désormais prises en charge par des acteurs qui sont presque totalement étrangers au monde des statistiques. Si bien que les statistiques universitaires ont beaucoup perdu en influence sur le développement de l'analyse de données :

> Aujourd'hui, la conception des systèmes est presque entièrement dans les mains des firmes qui développent des systèmes commerciaux pour l'analyse de données. Ce que les analystes de données utilisent en pratique est déterminé par ces firmes. Celles-ci travaillent dur pour produire les meilleurs systèmes possibles, mais l'innovation doit être limitée par leur besoin de rester bénéficiaires. Sans l'impulsion de la recherche consacrée à l'innovation, le progrès dans le traitement de données a été plus lent, et la créativité qui existe à l'intérieur des universités n'a eu pratiquement aucune influence. (Cleveland, 2001)

C'est pour surmonter cette crise professionnelle que plusieurs statisticiens américains jugent urgent d'en revenir aux propositions que formulait John Tukey dans les années 1960. William S. Cleveland – un autre chercheur statisticien des Bell Labs – joue un ici un rôle majeur en revalorisant le mouvement pour l'analyse de données au tout début des années 2000. Il propose l'expression « science des données » (*data sciences*) pour qualifier un domaine des statistiques qui serait à la fois plus étendu et redéfini (Cleveland, 2001).

Cette « science des données » est définie par plusieurs traits. Elle incite d'abord les statisticiens à s'engager dans des projets d'analyse de données en lien avec des chercheurs d'autres disciplines ou même d'entreprises. Plutôt que de définir les statistiques comme un ensemble d'outils, cette « science » vise à traiter des « problèmes » auxquels sont confrontés un grand nombre d'organisations et de mondes sociaux. Mais surtout, il s'agit de « faire la paix avec l'informatique », en faisant en sorte que les statisticiens s'engagent autant dans le développement d'outils informatiques que dans les modèles mathématiques. Le point majeur est alors, explique Cleveland, de fusionner les connaissances provenant des statistiques et celles issues de l'informatique :

> Les informaticiens n'ont qu'une connaissance limitée de la manière de penser l'analyse des données, tout comme les statisticiens n'ont qu'une connaissance limitée des environne-

> ments informatiques. Fusionner les bases de ces connaissances produirait une force puissante pour l'innovation. […] Cela veut dire que les départements de « science des données » devraient comporter des chercheurs qui consacrent leur carrière à faire avancer le traitement informatique de données et qui nouent des partenariats avec des informaticiens. (Cleveland, 2001)

Ainsi conçue, la « science des données » enjoint aux statisticiens de ne plus se définir professionnellement à partir des seules mathématiques, mais de s'intéresser tout autant à la conception d'outils informatiques et à la résolution de problèmes intéressant un grand nombre d'organisations situées en dehors des départements de statistiques. D'une nature profondément pluridisciplinaire, cette « science » se présente donc comme un assemblage de technologies et de savoirs situés au croisement des statistiques et de l'informatique.

Au tournant des années 2000, la « science des données » s'apparente donc à une proposition, formulée par des statisticiens américains, visant à résoudre la crise vécue par la profession. Elle se réclame alors explicitement du mouvement pour l'analyse de données, qui était apparu dans les années 1960. Cette « science des données » présente trois traits majeurs : (1) une étroite articulation entre des préoccupations scientifiques et industrielles ; (2) un intérêt pour des jeux de données volumineux et hétérogènes ; (3) une posture davantage centrée sur l'exploration des données que sur la confirmation de modèles statistiques.

2. Une « science » ouverte aux épistémologies des mondes sociaux

Depuis quelques années, un nombre croissant d'acteurs se revendique de cette « science des données », dont nous venons de voir qu'elle s'inscrit dans des dynamiques longues propres aux mondes statistiques et informatiques. Nous verrons à présent ce que cette extension doit à une ouverture assez large de cette « science » aux épistémologies des mondes sociaux auxquels elle s'adresse.

2.1. Quels sont les mondes sociaux concernés ?

Afin d'identifier quels sont les mondes sociaux aujourd'hui affectés par cet assemblage de technologies et de savoirs, nous avons procédé à une analyse du logiciel R. Créé à la fin des années 1990, R se présente à la fois comme un langage de programmation statistique et un logiciel de traitement de données. S'inscrivant dans la continuité du projet S conçu par les chercheurs des Bell Labs, R procède du mouvement pour l'analyse de données (De Leeuw, 2011). R a rapidement connu un succès considérable : d'abord utilisé par les statisticiens et les spécialistes de l'analyse informatique de données, il a ensuite très largement investi les mondes de l'enseignement des statistiques, de la recherche et de l'entreprise.

Projet *open source*, R évolue ainsi constamment à mesure que des contributeurs déposent de nouvelles briques logicielles – appelées *packages* – destinées à permettre des traite-

ments plus ou moins spécifiques. Au mois d'octobre 2014, le projet compte 5 873 *packages*, chacun d'entre eux étant désigné par un acronyme, associé à un responsable et accompagné d'un petit texte de présentation :

> Backtest
>
> Le *package* « Backtest » fournit des instruments pour explorer des hypothèses de portefeuilles concernant des instruments financiers (actions, obligations, échanges de créances, options financières, etc.)[8].

Nous avons lu l'ensemble des textes associés aux *packages*, en codant manuellement la référence éventuelle à des mondes sociaux situés en dehors des mondes statistiques et informatiques. Nous avons considéré que les mondes sociaux ciblés étaient un bon indicateur de la diffusion de cette « science des données » entendu comme assemblage de savoirs et de technologies[9].

Il en ressort qu'un quart des *packages* font explicitement référence à des mondes sociaux extérieurs aux statistiques et à l'informatique entendues comme domaines spécialisés. Comme l'indique le tableau 1 ci-dessous, ce sont des mondes sociaux très différents que ciblent explicitement ces briques logicielles.

Mondes sociaux	**Nombre de *packages***	**Pourcentage des *packages***
Génétique	379	26,5 %
Médecine, santé, épidémiologie, pharmacologie, toxicologie	163	11,4 %
Biologie	145	10,1 %
Écologie, biodiversité, éthologie, populations animales & végétales	119	8,3 %
Finance, assurance, analyse des risques	93	6,5 %
Sciences sociales, économie, démographie	80	5,6 %
Environnement, géologie, hydrologie, pédologie, océanographie	73	5,1 %
Sciences physiques, astronomies	51	3,6 %
Climat, météo, énergie et atmosphère	50	3,5 %
Psychologie, sciences cognitives, neurosciences	38	2,7 %

8. Jeff Enos, « The R project for statistical computing », http://www.r-project.org/, mis en ligne le 18 février 2010.
9. Notre analyse comporte deux limites. La première est que rien n'empêche un *package* ne s'adressant pas à un monde social précis d'être utilisé bien au-delà des mondes statistique et informatique. La seconde limite tient à ce que nous ne savons rien de l'utilisation effective des différents *packages* du projet. L'analyse que nous proposons ici témoigne néanmoins de l'inscription des briques logicielles dans une pluralité de mondes sociaux.

Services web	37	2,6 %
Chimie, biochimie	34	2,4 %
Géographie	24	1,7 %
Administrations & institutions statistiques	20	1,4 %
Marketing, benchmark, management	19	1,3 %
Sports, jeux et culture	17	1,2 %
Analyse du langage	16	1,1 %
Politique et politiques publiques	14	1,0 %
Sciences médico-légales, police, cryptographie	11	0,8 %
Agriculture, alimentation	10	0,7 %
Éducation & emploi	8	0,6 %
Médias & journalisme	8	0,6 %
Scientométrie, publications académiques	9	0,6 %
Aviation	4	0,3 %
Autres	9	0,6 %
Total	**1 431**	**100 %**

Tableau 1. Les mondes sociaux de R.

Plusieurs éléments apparaissent ici. Premièrement, on perçoit la très grande diversité des mondes sociaux auxquels s'adresse cet assemblage de savoirs et de technologies. De la santé à l'environnement, de la finance à la politique en passant par le marketing et le sport, rares sont les mondes qui ne sont pas mentionnés dans R. L'univers du religieux serait peut-être le seul à n'être pas explicitement visé. Deuxièmement, on remarque la très forte représentation des mondes scientifiques, certains d'entre eux occupant une place majeure – la génétique, la médecine et la biologie représentent à elles seules près de la moitié de l'ensemble des *packages* faisant explicitement référence à un monde social. La forte visibilité de la génétique s'explique par le fait que la recherche dans ce domaine consiste, depuis deux décennies, presque exclusivement à traiter des volumes considérables de données issues des procédés de séquençage (Keating et Cambrosio, 2012). D'autres domaines scientifiques accordent depuis plusieurs décennies une grande importance au traitement de données et à la modélisation – c'est le cas du climat, de l'économie ou de la physique (Armatte et Dahan, 2004) – si bien qu'il n'est pas surprenant de les trouver si bien représentés ici. Troisièmement, on constate une forte intrication des perspectives scientifiques et industrielles. Ceci est particulièrement visible dans le domaine des médicaments ou de l'analyse des risques, dans lesquels les sciences concernées sont étroitement associées à des finalités industrielles. Enfin, certains domaines sont plus éloignés de la recherche scientifique – le sport, la culture, les médias ou le journalisme – et on peut penser qu'il s'agit là de mondes plus récemment investis par l'analyse de données.

Le logiciel R constitue aujourd'hui l'une des infrastructures qui sous-tendent l'extension de la « science des données » à une grande diversité de mondes sociaux. Mais les algorithmes et les modèles statistiques qui constituent ces *packages* sont d'une grande hétérogénéité[10]. Parmi ceux qui sont les plus téléchargés[11], on trouve une diversité de modèles – régression, séries temporelles, analyse de réseaux, etc. –, de nombreux algorithmes de visualisation ainsi que des outils permettant d'interfacer R avec de nombreux logiciels d'analyse statistique et de gestion de bases de données.

2.2. L'importance des épistémologies locales

Lorsqu'on quitte le terrain des enjeux des *big data*, et que l'on se plonge dans les conseils pratiques qui accompagnent la formation des « scientifiques des données » (*data scientists*), on ne peut que constater la place importante accordée aux épistémologies locales. Ce phénomène s'inscrit dans la continuité de l'histoire du mouvement pour l'analyse de données, qui incitait le statisticien à s'intéresser aux problèmes rencontrés par les organisations situées en dehors des départements de statistiques.

Nous avons examiné plusieurs manuels du domaine. L'un d'entre eux s'intitule *Doing Data Science* et a été publié en 2013 par les éditions O'Reilly – qui s'adressent traditionnellement aux mondes de l'informatique et du Web. Ce manuel a été rédigé par deux statisticiennes qui ont travaillé dans le privé à l'issue de leur thèse (O'Neil et Schutt, 2013). Il s'appuie sur leur expérience d'enseignantes à l'université de Columbia. Leur cours s'adresse à des étudiants d'horizons divers – statisticiens, informaticiens, chercheurs en biomédecine, *marketers*, journalistes, etc. – intéressés par la perspective de « trouver des façons de résoudre des problèmes importants, qui ont souvent une valeur sociale, avec des données » (O'Neil et Schutt, 2013, p. 15).

D'emblée, l'ouvrage s'inscrit implicitement dans la filiation du mouvement pour l'analyse de données. Faire de la science de données, expliquent les auteures, consiste à s'intéresser à des jeux de données volumineux et hétérogènes (chiffres, données de capteurs, textes, données temporelles, etc.) et à concilier un intérêt pour les modèles statistiques avec un goût pour les dispositifs informatiques. Mais surtout, affirment-elles, il est indispensable de se fonder sur un ensemble de connaissances propres au monde social concerné. C'est ce que traduit le diagramme de Drew Conway qu'elles mettent en avant comme une représentation de ce qu'est la science des données :

10. L'analyse des descriptifs de chaque *package* ne permet pas de mesurer avec précision quels sont les éléments statistiques ou informatiques communs à des *packages* associés à des mondes sociaux différents. Lorsque les contributeurs décrivent les fonctionnalités de leur *package*, ils utilisent souvent des catégories soit très générales (*modeling*, *simulation*), soit très spécifiques au monde dans lequel ils s'inscrivent (*association analysis* ; *risk assessment*).

11. Nous nous appuyons ici sur la liste des 100 *packages* les plus téléchargés début 2013 (http://www.r-statistics.com/2013/06/top-100-r-packages-for-2013-jan-may/ ; dernière consultation le 8 mars 2017).

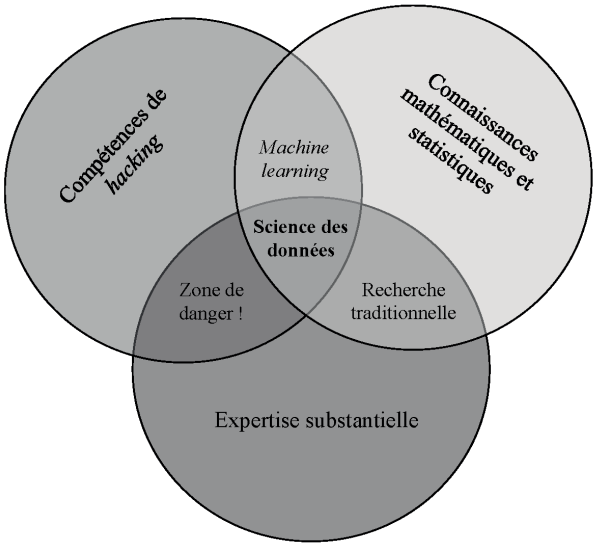

Figure 1. Diagrame de Venn de la science des données selon Drew Conway.

Source : Conway, 2013, adapté en français par les auteurs.

Ce diagramme enjoint de prendre appui à la fois sur les savoirs statistiques, les compétences informatiques et l'expertise propre à un domaine précis. Autrement dit, il n'est pas question de mobiliser des outils informatiques associés à des modèles statistiques sans égard pour les savoirs propres au monde social auquel le *data scientist* s'intéresse.

Mais l'ouverture aux épistémologies locales ne doit pas se limiter pas à une seule incantation. Elle imprègne au contraire l'ensemble des étapes du « processus de la science des données » qui sont décrites dans le manuel, et synthétisées dans le graphique ci-dessous.

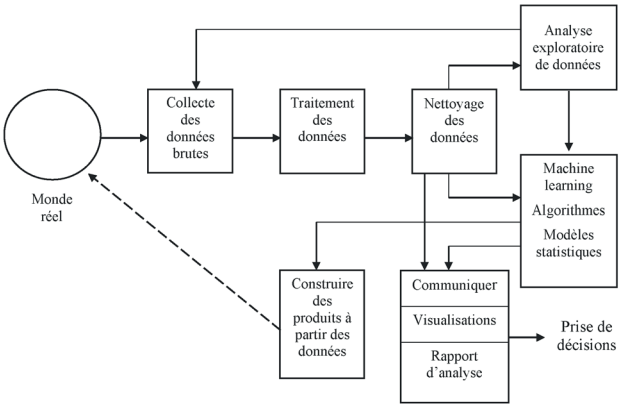

Figure 2. Le processus de la science des données.

Source : O'Neil et Schutt, 2013, p. 41, adapté en français par les auteurs.

L'« analyse exploratoire des données » est ici présentée comme une étape centrale du processus. Explicitement inscrite dans l'héritage de John Tukey, cette opération organise les interactions entre le *data scientist* et les spécialistes du monde considéré (ingénieurs, responsables, etc.). Elle assure l'ouverture du processus aux épistémologies locales, c'est-à-dire une sensibilité à la fois aux questions et aux problèmes qui sont jugés importants par les acteurs des mondes sociaux considérés.

En pratique, les auteures expliquent que l'analyse exploratoire doit intervenir une fois les premiers jeux de données constitués. À cette étape, le *data scientist* n'a aucun modèle statistique à l'esprit, ni ne sait précisément à quelle question ou problème il va se consacrer. Il se met d'abord à parcourir les données de façon systématique – au moyen de distributions et de graphiques – de façon à se familiariser avec ces données. Identifiant les données manquantes et le degré de « propreté » des données, il comprend alors de quelle façon les données ont été construites et de quelle manière le dispositif de collecte doit éventuellement être modifié. Mais surtout, cette première exploration permet au *data scientist* d'échanger, *via* des représentations graphiques, avec des acteurs qui n'ont pas l'habitude d'analyser des données. C'est au terme de cette exploration et de ces interactions que le *data scientist* identifie progressivement à la fois le problème à résoudre, le produit susceptible d'être conçu, et les algorithmes et modèles statistiques qu'il faut concevoir ou ajuster.

Nous sommes ici bien éloignés des discours qui accompagnent souvent le « Big Data ». Non seulement ces spécialistes reconnaissent le caractère construit des données[12] et les choix qui sous-tendent la sélection des modèles, mais ils aussi mettent en avant la nécessité d'intégrer les épistémologies locales dans la mise en œuvre de cette « science des données ». C'est cette plasticité qui explique en grande partie, croyons-nous, la faculté de cette science à se diffuser dans les mondes sociaux les plus divers.

3. La science des données à l'épreuve des mondes sociaux

Si la plasticité épistémologique de cette « science des données » est une condition importante de sa diffusion, c'est que la plupart des mondes sociaux ne lui réservent pas d'emblée un accueil bienveillant. Ces mondes sociaux soumettent en effet cette « science » à un ensemble d'épreuves (Lemieux, 2007). C'est ce que nous allons maintenant examiner dans le cas de deux mondes sociaux dans lesquels nous enquêtons depuis plusieurs années : le monde du journalisme et celui de la santé. Comme nous l'avons vu précédemment, ces deux mondes sont dans une position opposée par rapport à la science des données (cf. tableau 1). Alors que le monde de la santé y occupe une place de choix depuis de nombreuses années, celui du journalisme ne s'y est ouvert que plus récemment.

12. Contrairement à ce que pourrait laisser penser le schéma, les auteures précisent que la collecte des données repose sur un ensemble de choix : « Vous, le *data scientist*, êtes l'observateur. Vous transformez le monde en données, et ceci est un processus totalement subjectif, en aucun cas objectif. » (O'Neil et Schutt, 2013, p. 19).

3.1. Le « journalisme de données » est-il un journalisme ?

Dans le monde journalistique américain, il existe une tradition bien établie qui valorise la mobilisation, par les reporters eux-mêmes, des statistiques et des logiciels d'analyse de données. Dès la fin des années 1960, des reporters ont collaboré avec des chercheurs en sciences sociales et entrepris d'analyser eux-mêmes des données dans le cadre de leurs investigations. Le courant du *computer-assisted reporting* a ensuite participé à la définition de normes précisant les usages légitimes de l'analyse de données du point de vue de l'intérieur du monde journalistique américain (Parasie et Dagiral, 2013).

Dans la seconde moitié des années 2000, les usages journalistiques du traitement de données ont connu un essor important, bien au-delà du territoire américain. Des individus issus des mondes informatiques et statistiques (développeurs web, activistes *open data*, analystes de données) ont été employés par des organisations de presse ou ont fondé leurs propres *start-up* à prétention journalistique. Les réalisations journalistiques ont pris des formes très diverses – visualisations, applications, bases de données en ligne, articles textuels, etc. – et se sont appuyées sur des algorithmes très variés – traitement du langage, visualisation, *machine learning*, etc.

Des échanges plus ou moins vifs ont eu lieu à l'intérieur des rédactions dans lesquelles ces individus sont intervenus. Fondamentalement, les journalistes se posent la question suivante : à quelles conditions la mobilisation des algorithmes et des bases de données permet-elle de produire une information qui soit valide d'un point de vue journalistique ? Nous avons pu analyser, dans nos recherches, les tensions épistémologiques qui sont aujourd'hui à l'œuvre à l'intérieur comme à l'extérieur de certaines rédactions américaines (cf. Parasie, 2013 et 2014). Ce sont souvent les journalistes formés au sein du courant du *computer-assisted reporting* qui interrogent la valeur journalistique de ces nouveaux produits proposés au public.

Cette mise à l'épreuve a notamment concerné deux points majeurs. Le premier a trait à la qualité des données traitées dans le cadre de ces projets. Un grand nombre de journalistes rompus à l'investigation ont mis en avant la nécessité de vérifier les données, et de satisfaire ainsi une norme majeure de leur univers professionnel. En 2009, des journalistes du *Los Angeles Times* ont ainsi accusé les responsables du site EveryBlock.com – un site d'information locale traitant un grand nombre de données provenant notamment de la police – de diffuser de fausses informations concernant la criminalité dans un quartier de la métropole californienne. Mais, dès lors que les bases de données exploitées sont d'une taille importante, la vérification systématique des données peut rapidement se révéler impossible à effectuer. Nous avons pu observer que des négociations avaient eu lieu dans les rédactions, et que plusieurs compromis avaient pu être trouvés localement. Les journalistes peuvent ainsi décider de procéder à la vérification systématique d'un échantillon des données ; ils peuvent aussi mettre en garde leurs lecteurs contre une possible imprécision des données ; ils peuvent encore concevoir des algorithmes qui intègrent le risque d'erreur associé à la précision des données (Parasie, 2014).

Le second point concerne l'importance qu'accorde le « journaliste de données » à l'expertise des acteurs qui connaissent le sujet dont il est question. Plusieurs journalistes ont ainsi critiqué les projets d'analyse de données qui abordent un problème sans tenir compte des acteurs sociaux ou des chercheurs qui le connaissent intimement. Nate Silver, qui a animé

le blog Fivethirtyeight du *New York Times* – consacré à l'analyse de données électorales notamment – a ainsi été pris à parti à l'intérieur du monde journalistique :

> Vous ne pouvez pas extraire le sens des données à partir d'un modèle unique. Peu importe que vous soyez très fort pour analyser des choses avec le langage statistique R, vous aurez des problèmes si vous n'avez pas une compréhension profonde de l'origine des données, de la façon dont elles ont été collectées, filtrées et traitées, de leurs forces et de leurs faiblesses. […] Il s'agit aussi d'enquêter en collaboration avec des chercheurs, qui sont les seuls à vraiment connaître les données. Les médias qui pratiquent le journalisme de données et d'investigation le plus solide aujourd'hui, comme ProPublica, travaillent ainsi régulièrement. Leurs journalistes sont très conscients des limites de leur savoir. (Cairo, 2014)

Si l'on s'en tient au cas américain, on mesure donc à quel point le monde journalistique est vigilant quant à la valeur journalistique des nouveaux produits ainsi offerts au public.

3.2. Limiter l'analyse des données à l'échelle d'un l'individu ? Le *quantified self*

L'émergence du mouvement du *quantified self* (« quantification de soi », abrégé en QS) et des pratiques qui lui sont associées constitue un second exemple récent et original des tensions qui se font jour dans les manières d'analyser des données numériques. Celles-ci sont en l'occurrence produites par des capteurs et des dispositifs variés souvent portés à même le corps (bracelets, téléphones mobiles, etc.). Les données que ces capteurs et leurs algorithmes génèrent visent par exemple à comptabiliser un nombre de pas effectués, à proposer une représentation graphique des cycles du sommeil d'un individu, ou encore à mesurer les variations de son rythme cardiaque, pour n'en retenir que quelques-unes. Les *data products* vendus par les entreprises qui commercialisent ces objets (ou encore par des services web intermédiaires spécialisés) fournissent ainsi à leurs utilisateurs un mélange de chiffres, de visualisations et de tableaux de données permettant la comparaison et le suivi dans le temps. Selon les cas, ils rendent ou non possibles l'extraction des données enregistrées et à partir de là un travail de ré-analyse de ces données hors du cadre du service originel. Pour ces entreprises, l'analyse de données est donc centrale à plusieurs titres : le service fourni au client est *in fine* constitué par la présentation des données et leurs formes d'analyse ; et la pertinence même du dispositif et du service est sous-tendue par les modèles de données utilisés, eux-mêmes alimentés par les données fournies par les usagers équipés.

En questionnant les frontières entre les catégories de santé, de bien-être et d'activité physique ou cognitive, le QS s'appuie donc largement sur la science des données[13] et ses outils autant qu'il interroge les pratiques de mesure et d'analyse statistique en vigueur au sein des sciences de la vie et de la santé. Si l'analyse des données de population, et, plus largement, le domaine de l'épidémiologie, manifeste un intérêt certain pour les objets connectés, les

13. Parmi les entretiens réalisés à San Francisco en septembre 2012 auprès de plusieurs *start-up* du domaine (QS), trois l'ont été auprès d'acteurs s'auto-désignant comme « *data scientists* ». Tous avaient une expérience préalable au sein d'entreprises dédiées à la conception de services web, sans lien apparent avec des questions de santé et de données numériques liées au corps ou à la génétique, avant de s'orienter vers ce domaine émergent.

controverses sont vives quant aux pratiques d'analyses statistiques effectivement accomplies au sein de ces mouvements. Celles-ci sont en effet très discutées et souvent critiquées, à la fois du point de vue de la qualité des mesures réalisées, des protocoles d'analyse statistiques mis en œuvre et des finalités éthiques relatives aux résultats et aux services produits (Neff, 2013 ; Barret *et al.*, 2013). En son sein, pourtant, ce mouvement compte nombre de chercheurs et d'acteurs du domaine de la santé publique, mais aussi de patients intéressés par les perspectives offertes par ces technologies.

Les premières enquêtes de sciences sociales dont nous disposons sur les acteurs et premiers usagers des dispositifs de quantification de soi proposent des descriptions très contrastées des pratiques d'analyse de données observées. Des auteurs soulignent combien les données fournies, telles quelles, peuvent ne pas faire sens, et ne font que très exceptionnellement l'objet d'une analyse originale par les usagers eux-mêmes (Pharabod *et al.*, 2013) ; ou que, lorsqu'elles sont analysées par les usagers, elles requièrent un engagement et un travail conséquent (Ruckenstein, 2014). S'intéressant pour leur part à des passionnés des *meetups* dédiés à ces expérimentations, d'autres enquêtes pointent les tensions entre l'analyse des mesures individuelles et l'analyse des données à des échelles collectives : puisque, lors des présentations individuelles, « n = 1 » prévaut (quitte à analyser des volumes considérables de données pour cet individu), il est possible d'y voir une « pratique alternative du Big Data » :

> Le Big Data n'a pas à voir qu'avec les grandes institutions ; c'est aussi une question de subjectivités. (Nafus et Sherman, 2014, p. 1786)

Cette tension essentielle constitue donc une épreuve d'échelle. D'un côté, l'analyse de données est fondamentalement ajustée à l'individu (une autre désignation du travail d'analyse de ces données, à savoir « *personal analytics* », insiste d'ailleurs sur leur dimension personnelle autant que sur l'ubiquité de leur nature) ; de l'autre, les données personnelles produites représentent une perspective de connaissance prometteuse à des échelles collectives, dont l'épidémiologie est plus familière. Pour autant, de nombreux promoteurs du QS opposent une fin de non-recevoir à tout ou partie des chercheurs intéressés par les opportunités de ré-analyse des données collectées par ces premiers utilisateurs[14] – bien que, dans le même temps, de nombreux praticiens mettent en ligne sans restriction aucune l'ensemble des données les concernant, pour l'usage des communautés d'intérêt qui leur sont proches (Dagiral, 2014).

14. Nous avons ainsi pu observer, lors du second congrès européen du QS organisé à Amsterdam en mai 2013, l'opposition d'une partie des participants à la proposition de fonder une revue du QS articulée autour des expérimentations de ses participants et de l'analyse des données collectées, imaginée par d'autres membres – pour partie chercheurs eux-mêmes – selon le format d'une revue scientifique. Ce projet n'a donc pas vu le jour.

Conclusion

Au terme de ce chapitre, nous espérons avoir convaincu le lecteur de l'intérêt sociologique qu'il y a à aller au-delà de la seule discussion sur le Big Data et ses enjeux. En faisant porter l'analyse sur des pratiques, des savoirs, des individus et des technologies situées – ce à quoi la notion de « science des données » permet plus sûrement d'accéder –, on se donne les moyens de décrire et d'expliquer l'extension de l'analyse de données à une grande diversité de mondes sociaux. Les travaux d'Alain Desrosières invitaient déjà à une telle perspective, en mettant en lumière le rôle majeur des dynamiques internes aux mondes statistiques (Desrosières, 2008 ; 2014).

Cette perspective permet de mieux comprendre sur quoi repose la diffusion des pratiques d'analyse de données dans une diversité de mondes sociaux. Contrairement à ce que mettent en avant les promoteurs du Big Data, cette diffusion ne s'explique pas par la transformation profonde des formes de connaissance qui ont cours dans chacun de ces mondes sociaux. Nous avons plutôt vu à quel point cette science des données fait la part belle aux façons de connaître qui dominent les mondes sociaux. C'est même précisément la plasticité de cet assemblage de savoirs, de pratiques et de technologies, et son ouverture aux façons de connaître locales, qui expliquent cette diffusion.

L'approche que nous avons esquissée ici ouvre plusieurs chantiers. Le premier est historique. Parce que la plupart des travaux historiques ont privilégié l'étude des activités statistiques au sein des institutions publiques, on sait peu de choses sur la manière dont l'analyse de données a été prise en charge dans le monde de l'entreprise ces dernières décennies. Le deuxième chantier concerne le rôle des logiciels d'analyse statistique – dont l'histoire demeure elle aussi méconnue – et des infrastructures dans la diffusion de cette science des données. Nous avons pointé ici le rôle du langage R, mais d'autres travaux seraient nécessaires pour saisir de quelle manière ce dispositif alimente l'essor de l'analyse de données. Enfin, le dernier chantier concerne les épreuves auxquelles est soumise cette science des données dans un grand nombre de mondes sociaux. Il est particulièrement crucial de saisir de quelle manière ces savoirs et ces technologies d'analyse de données sont contestés, traduits et transformés.

Références

Anderson C. (2008), « The end of theory: The data deluge makes the scientific method obsolete », *Wired*, 28 juin 2008, https://www.wired.com/2008/06/pb-theory/ (consulté le 8 mars 2017).

Armatte M. et Dahan Dalmedico A. (2004), « Modèles et modélisations, 1950-2000 : nouvelles pratiques, nouveaux enjeux », *Revue d'histoire des sciences*, vol. 57, nº 2, p. 243-303.

Barrett M.A. Humblet O., Hiatt R. et Adler N.E. (2013), « Big data and disease prevention: From quantified self to quantified self communities », *Big Data*, vol. 1, nº 3, p. 168-175.

Boyd D. et Crawford K. (2012), « Critical questions for big data », *Information, Communication & Society*, vol. 15, nº 5, p. 662-679.

Cairo A. (2014), « Data journalism needs to up its own standards », *niemanlab.org*, http://www.niemanlab.org/2014/07/alberto-cairo-data-journalism-needs-to-up-its-own-standards/ (publié le 9 juillet 2014, consulté en novembre 2014).

Ceruzzi P. (1993), « Une révolution inattendue. Les premiers pas de l'informatique (1935-1985) », *Culture technique*, nº 28, p. 164-177.

Chambers J. (1999), « Computing with Data: Concepts and Challenges », *The American Statistician*, vol. 53, nº 1, p. 73-84.

Cleveland W.S. (2001), « Data science: An action plan for expanding the technical areas of the field of statistics », *International Statistical Review*, vol. 69, nº 1, p. 21-26.

Conway D. (2013), « The data science Venn diagram », http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram (publié en ligne le 26 mars 2013, consulté en novembre 2014).

Dagiral E. (2014), « Step-by-step self-learning? The quantification and interpretation of walking activities », communication au congrès de l'International Sociological Association, Yokohama, 16 juillet 2014.

De Leeuw J. (2011), « Statistical software: An overview », *in* Lovric M. (dir.), *International Encyclopedia of Statistical Science*, Springer, p. 1470-1473.

Desrosières A. (2008), *Gouverner par les nombres. L'argument statistique II*, Paris, Presses des Mines.

Desrosières A. (2014), *Prouver et gouverner. Une analyse politique des statistiques publiques*, Paris, La Découverte.

Friedman J.H. (1997), « Data mining and statistics: What's the connection? », *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*.

Friedman J.H. (2001), « The role of statistics in the data revolution? », *International Statistical Review*, vol. 69, nº 1, p. 5-10.

Galison P. (1996), « Computer simulation and the trading zone », *in* Galison P. et Stump D.J. (dir.), *The Disunity of Science. Boundaries, Contexts and Power*, Stanford, Stanford University Press, p. 118-157.

Gitelman L. (dir) (2013), *« Raw Data » is an Oxymoron*, Cambridge, MIT Press, coll. « Infrastructure studies ».

Grier D.A. (2005), *When Computers* Were *Humans*, Princeton, Princeton University Press.

Keating K. et Cambrosio A. (2012), « Too many numbers: Microarrays in clinical cancer research », *Studies in History and Philosophy of biological and biomedical sciences*, vol. 43, p. 37-51.

Knorr-Cetina K. (1999), *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge (Mass.), Harvard University Press.

Lemieux C. (2007), « À quoi sert l'analyse des controverses ? », *Mil neuf cent*, vol. 1, nº 25, p. 191-212.

Malinvaud E. (1970), « L'analyse des données », *Annales de l'INSEE*, nº 4, p. 3-8.

Mayer-Schönberger V. et Cukier K. (2013), *Big Data. A Revolution that Will Transform how we Live, Work and Think*, Boston et New York, Houghton Mifflin Harcourt.

McGrayne S.B. (2011), *The Theory That Would Not Die. How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant From Two Centuries of Controversy*, New Haven et Londres, Yale University Press.

Nafus D. et Sherman J. (2014), « This one does not go up to 11: The quantified self movement as an alternative big data practice », *International Journal of Communication*, vol. 8, p. 1784-1794.

Neff G. (2013), « Why big data won't cure us », *Big data*, vol. 1, nº 3, p. 117-123.

O'Neil C. et Schutt R. (2013), *Doing Data Science. Straight Talk from the Frontline*, Sebastopol, O'Reilly.

Parasie S. (2013), « Des machines à scandale. Éléments pour une sociologie morale des bases de données », *Réseaux*, nº 178-179, p. 127-161 (en ligne : https://www.cairn.info/revue-reseaux-2013-2-page-127.htm).

Parasie S. (2014), « Data-driven revelation ? Epistemological tensions in investigative journalism in the age of "big data" », *Digital Journalism*, DOI : 10.1080/21670811.2014.976408 (publié en ligne le 19 novembre 2014).

Parasie S. et Dagiral É. (2013), « Data-driven journalism and the public good. Computer-assisted reporters and programmer-journalists in Chicago » *New Media and Society*, vol. 15, nº 6, p. 853-871.

Pharabod A.-S., Nikolski V. et Granjon F. (2013), « La mise en chiffres de soi. Une approche compréhensive des mesures personnelles », *Réseaux*, nº 177, p. 97-129 (en ligne : https://www.cairn.info/revue-reseaux-2013-1-page-97.htm).

Ruckenstein M. (2014), « Visualized and interacted life: Personal analytics and engagements with data doubles », *Societies*, vol. 4, nº 1, p. 68-84 (en ligne : http://www.mdpi.com/2075-4698/4/1/68).

Smith D. (2013), « R users: Be counted in Rexer's 2013 Data Miner Survey », *Revolution Analytics Blog*, 30 janvier 2013, http://blog.revolutionanalytics.com/2013/01/r-users-be-counted-in-rexers-2013-data-miner-survey.html (consulté le 13 novembre 2014).

Star S. et Griesemer J. (1989), « Institutional ecology, "translation"' and boundary objects: Amateurs and professionals in Berkeley's Museum of verterbrate zoology », *Social Studies of Science*, vol. 19, nº 3, p. 387-420.

Strasser B.J. (2012), « Data-driven sciences: from wonder cabinets to electronic databases », *Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 43, nº 1, p. 85-87.

Strauss A. (1992), *La Trame de la négociation. Sociologie qualitative et interactionnisme*, Paris, L'Harmattan.
Tukey J.W. (1962), « The future of data analysis », *The Annals of Mathematical Statistics*, vol. 33, n° 1, p. 1-67.
Tukey J.W. (1977), *Exploratory Data Analysis*, Reading (Mass.), Addison-Wesley.
Tukey J.W. (1980), « We need both exploratory and confirmatory », *The American Statistician*, vol. 34, n° 1, p. 23-25.

Infrastructures de données bibliométriques et marché de l'évaluation scientifique

David Pontille

Sociologue et directeur de recherche au CNRS, membre du Centre de sociologie de l'innovation (Mines ParisTech, PSL Research University, CSI-i3 UMR CNRS 9217)

Didier Torny

Directeur de recherche au CNRS et membre du Centre de sociologie de l'innovation (Mines ParisTech, PSL Research University, CSI-i3 UMR CNRS 9217)

Les infrastructures qui soutiennent l'évaluation quantitative des productions scientifiques génèrent un volume de données en forte augmentation depuis les années 2000. Cette masse d'informations n'est pas uniquement due à l'entrée en scène de différents producteurs de données et l'avènement de l'Internet à haut débit. Elle se caractérise aussi par une accessibilité grandissante qui accentue considérablement les formes de calculabilité : des données de diverses natures sont sujettes à de nombreux calculs secondaires, et associées à des formes de visualisation inédites. Alors que, jusque dans les années 1980, elles étaient réservées à de petits cercles de spécialistes ayant négocié l'accès à des données propriétaires, les analyses bibliométriques sont aujourd'hui disponibles et parfois gratuites pour des chercheurs ordinaires.

Pour autant, ces infrastructures demeurent difficilement saisissables si on les réduit à un épiphénomène des « *big data* ». Non seulement ce terme masque la concurrence particulièrement vive qui oppose actuellement différentes conceptions de l'excellence scientifique et du cercle des pairs, mais il amalgame également une multiplicité d'éléments qui sont agencés dans les technologies de calcul en plein essor depuis le milieu des années 1950. Faire un pas de côté vis-à-vis d'une telle approche englobante permet de saisir comment les opérations de dénombrement, d'indexation et de classement reposent sur un fin « travail d'infrastructure » (Bowker et Star, 1999) où innovations techniques, appariements industriels et décisions politiques sont intimement mêlées. Cette posture ouvre sur la grande hétérogénéité des infrastructures informationnelles dédiées à l'évaluation quantitative des productions scientifiques. Partant, nous prendrons soin ici de distinguer trois principaux composants[1].

Tout d'abord, les *algorithmes* sont des formules mathématiques, développées à partir des années 1940, sous la forme de simples dénombrements, de ratios entre deux quantités ou, parfois, de calculs algébriques plus élaborés. Ensuite, les *jeux de données* constituent des sources organisées de documents (bases de résumés, d'articles, de revues, d'ouvrages, d'auteurs, etc.) mis en séries et classés selon différents critères. Ils peuvent être construits expressément, à titre individuel ou collectif, pour les besoins d'une recherche, mais ils font

1. Ce chapitre est une version remaniée de Pontille et Torny (2013).

aussi l'objet d'investissements lourds de la part d'organismes publics ou d'entreprises privées. Enfin, les *outils bibliométriques* associent de manière pérenne certains algorithmes à des jeux de données spécifiques. Ils reposent sur des combinaisons plus ou moins complexes entre formules et données, et leur degré d'ouverture est variable (accès aux résultats, aux données du calcul, protection intellectuelle des outils).

Une telle distinction nous permettra de repérer différentes configurations qui sont définies par des agencements bien particuliers entre jeux de données, algorithmes et outils bibliométriques. Cette approche en termes de configurations s'avère pertinente pour caractériser l'évolution contemporaine des pratiques d'évaluation quantitative des productions scientifiques, et notamment l'émergence de technologies de calcul qui élargissent l'analyse auparavant limitée aux citations dans les articles de revues vers des mesures d'audience et d'attention.

Nous commencerons par rappeler les principales conditions d'émergence, déjà largement documentées, d'une première configuration au milieu des années 1960, caractérisée par la position monopolistique d'un unique producteur de jeux de données, l'Institute for Scientific Information (ISI), proposant une palette d'outils bibliométriques. Nous examinerons ensuite une deuxième configuration qui se dessine au milieu des années 2000 avec la multiplication des acteurs proposant des technologies de calcul et renouvelant l'offre bibliométrique. Nous identifierons enfin des outils de mesure alternatifs qui jettent potentiellement les bases d'une troisième configuration, ouvrant considérablement l'horizon des formes d'évaluation des productions scientifiques.

1. L'Institute for Scientific Information : au centre de l'analyse des citations

En 2009, le rédacteur en chef du *Journal of Sleep Research* commençait ainsi un éditorial, intitulé « The race for the impact factor », où il décrivait l'attente générée par la parution annuelle du *Journal of Citation Reports* (JCR) qui compile le facteur d'impact des revues :

> Tous les ans, au mois de juin, les rédacteurs en chef et les éditeurs deviennent nerveux et agités, et même peut-être insomniaques, en attendant impatiemment le facteur d'impact de leur revue (IF) de l'année courante, calculé par Thomson Reuters (auparavant Thomson Scientific). Ces dernières années, l'IF a acquis une influence mythique sur le prestige des revues et des articles qu'elles publient, ce qui a conduit à une « course pour le facteur d'impact ». Les rédacteurs en chef ont essayé d'augmenter l'IF de leur revue autant qu'il est possible, et les auteurs ont essayé de publier leurs manuscrits dans des revues avec le plus haut IF possible[2].(Lavie, 2009, p. 283)

Dans l'ensemble des sciences biomédicales et dans une partie des sciences physiques et des sciences sociales, les commentaires sur les résultats du JCR sont devenus monnaie courante. Aujourd'hui vendu comme « un moyen systématique et objectif d'évaluer, de manière critique, les meilleures revues mondiales, à l'aide d'informations quantitatives, statistiques, fondées

2. Sauf mention contraire, les citations sont traduites en français par les auteurs du chapitre [NdE].

sur des données de citation[3] », le JCR est pourtant le fruit d'une trajectoire faite d'incertitudes et d'infléchissements inattendus. Conçu dans une tradition des sciences de la documentation, il était initialement destiné aux bibliothécaires des universités américaines afin qu'ils disposent d'un moyen objectif de gestion des collections de périodiques et de sélection des revues à acheter et à conserver. Outil bibliométrique servant à repérer l'information pertinente dans la masse des publications disponibles, le JCR est rapidement devenu un instrument de recherches bibliométriques, ainsi qu'un ingrédient clé de certaines politiques d'évaluation des universités et des chercheurs. Trois usages très différents qui, contrairement aux récits dominants sur le sujet, étaient d'emblée considérés par son fondateur Eugene Garfield :

> En poursuivant leur analyse, les bibliothécaires et chercheurs en information scientifique peuvent constituer des volumes d'articles fréquemment demandés. Cela semblerait stupide d'envoyer des volumes de revues qui ne seraient empruntés que pour un petit nombre d'articles […] Parallèlement, la même information pourrait être prudemment utilisée à des fins de sélection et d'évaluation (par exemple par les comités Nobel). J'ajouterais, cependant, que de nombreuses personnes sont intéressées par les index de citation, qui, évidemment, faciliteront l'évaluation d'articles clés. En effet, si j'en avais le temps, je démontrerais comment on peut utiliser les réseaux de citations afin de produire des études historiques et sociologiques extrêmement intéressantes. (Garfield et Sher, 1963, p. 200-201)

Afin que de tels usages puissent se développer, d'autres éléments devaient être stabilisés avant même la première publication du JCR en 1975. Loin d'être un document isolé, le JCR fait en effet partie d'une ligne de produits développés par l'ISI. Parmi ceux-ci, des jeux de données ont rapidement acquis de l'importance, à l'instar des *Current Contents* qui consistaient en une édition papier hebdomadaire des sommaires de plusieurs centaines de revues quelques semaines avant leur parution. Et c'est en partie à l'aide des profits dégagés par cette publication payante que Garfield a pu financer, non sans difficultés, la production d'autres outils devenus célèbres : le *Journal Impact Factor* (JIF) et le *Science Citation Index* (SCI).

1.1. Des jeux de données multidisciplinaires

Comme l'ont montré Archambault et Larivière (2009), le JIF, stabilisé en tant qu'algorithme en 1963, articule plusieurs techniques bibliométriques : le nombre total de citations élaboré par Gross et Gross (1927) afin de mesurer l'influence d'une revue au sein d'un domaine et de repérer ainsi un petit nombre de « revues centrales » dans la littérature scientifique (Bradford, 1934) ; l'usage d'un ratio, proposé par Raisig (1960), pour éliminer l'effet de structure engendré par le nombre total de citations d'une revue, fortement corrélé au nombre total d'articles qu'elle publie ; la période de référence de deux ans, définie par Martyn et Gilchrist (1968), pour calculer les citations reçues par une revue durant l'année suivante. L'agencement de ces caractéristiques donne au JIF son caractère novateur. Si la quantité de citations marque

3. https://web.archive.org/web/20140917021216/http://thomsonreuters.com/journal-citation-reports/ (tous les liens cités dans ce chapitre ont été consultés le 17 septembre 2014).

l'influence, c'est bien celle de la revue en tant qu'unité et non comme somme d'articles. Dans la perspective du JIF, la revue est l'entité dont on mesure l'impact, quels que soient les écarts entre les différents articles qui y sont publiés.

De son côté, le jeu de données rassemblant les revues sélectionnées devait être élaboré de toutes pièces. Wouters (1999) a bien montré que l'ambition première de Garfield était de réaliser un index de citations sur le modèle du *Shepard's Citator*, une base de données couvrant l'ensemble des décisions judiciaires américaines et permettant de savoir si une jurisprudence est encore valable. Le SCI n'a cependant vu le jour qu'avec l'appui motivé de plusieurs personnes-relais, notamment les généticiens Gordon Allen et Joshua Lederberg, auprès d'institutions scientifiques majeures telles que les National Institutes of Health et la National Science Foundation. Ces divers soutiens ont permis à Garfield de placer son projet au centre d'intenses débats sur la politique de la recherche et sur l'information scientifique au début des années 1960 aux États-Unis.

Le SCI est paru en 1964 dans une version minimale, son élaboration nécessitant un travail manuel fastidieux d'examen des revues et d'enregistrement des citations. Sa grande originalité à l'époque résulte de cette construction systématique : couvrir un vaste ensemble de domaines scientifiques. En renonçant aux découpages disciplinaires, ce jeu de données permettait de signaler des références pertinentes en circulant dans des domaines parfois inattendus. Imaginé comme un outil sophistiqué d'exploration de la littérature scientifique, le SCI rencontrait également les préoccupations d'une « science de la science » prônée par de Solla Price (1963), qui trouvait ainsi des prolongements aussi bien en sociologie des sciences qu'en scientométrie.

Dans une logique d'extension, suscitée par cet intérêt, l'ISI créait en 1972 un nouveau jeu de données, le *Social Sciences Citation Index* (SSCI). La liste des revues de sciences sociales incluses résulte d'une sélection à partir de trois sources : celles fortement citées dans le SCI, celles distinguées par la littérature en sciences de l'information et celles désignées par des chercheurs directement sollicités. Poursuivant son processus d'extension, l'ISI rendait public un autre jeu de données en 1978, *Arts & Humanities Citation Index*, qui a nécessité une adaptation de la nature des citations retenues aux spécificités de ces domaines. Ces différents jeux de données couvraient ainsi l'ensemble des disciplines à partir d'une sélection raisonnée, à l'aide du JIF, de revues supposément centrales à la suite du processus d'étalonnage. Autrement dit, l'élaboration de jeux de données multidisciplinaires n'était en aucun cas voué à accueillir l'ensemble des revues disponibles.

1.2. Le succès du Journal of Citation Reports

L'apparition du JCR en 1975 a marqué un tournant. Cet outil bibliométrique regroupe, sur les jeux de données produits par l'ISI, les résultats de plusieurs algorithmes (nombre total de citations, JIF et « *immediacy index* »). Avec sa parution annuelle, il délivre ainsi des mesures régulières portant directement sur les revues.

Cette nouvelle offre a donné lieu à des usages imprévus de la part de différents acteurs. Tout d'abord, les revues spécialisées dans les articles de synthèse (*review articles*), qui jusque-là

étaient peu considérées, s'en sont emparées et ont publié des éditoriaux d'autosatisfaction pour souligner leur grande influence :

> Une analyse récente […] de 50 revues de botanique a révélé que, pour le nombre d'articles cités, l'*Annual Review of Phytopathology* […] se classait seconde […], en dépit […] d'un taux d'autocitation extrêmement bas. Cela en dit beaucoup sur l'importance et le rôle des articles de synthèse. (*Annual Review of Phytopathology,* 1975)

Ensuite, des chercheurs ont utilisé les résultats du JCR pour dessiner de nouvelles hiérarchies internes à leur discipline. Par rapport à des classements construits, au début des années 1970, sur les jugements agrégés des pairs (Pontille et Torny, 2010), ceux élaborés à partir du JIF recueillaient de nombreux soutiens, notamment pour dénoncer l'injustice de tel ou tel processus de recrutement (Fabbri, 1987 ; Aiuti *et al.*, 1991). D'autres chercheurs faisaient un pas supplémentaire en modifiant leurs pratiques de publication et de citation afin d'être bien positionnés dans les hiérarchies inédites favorisées par cet outil bibliométrique.

Enfin, cette nouvelle centralité du JCR est devenue visible à travers les critiques récurrentes concernant son pouvoir, dont une première version élaborée apparaît en 1993 sous la plume d'un éditeur de *Science* prenant l'exemple d'un directeur de laboratoire québécois qui « répartissait les ressources et promotions dans son laboratoire sur la base d'un système de notes dans lequel […] le nombre de citations pesait pour 40 % » (Taubes 1993, p. 885). Par la suite, des critiques ont pointé les lacunes engendrées par le centrage anglo-saxon des jeux de données (Hicks, 1999), ainsi que les limites découlant de leur fabrication et de leur privatisation (Weingart, 2005 ; Rossner *et al.*, 2007).

Le JCR s'installait progressivement au sein du monde académique. Ce mouvement s'est accéléré avec la mise en ligne des services proposés par l'ISI, notamment l'élaboration du *Web of Science* (WoS), un dispositif permettant d'interroger le JCR à partir de n'importe quelle bibliothèque universitaire possédant un abonnement. Au début des années 2000, l'ISI était donc en situation de quasi-monopole pour la production de jeux de données à destination d'outils bibliométriques[4]. Parmi la large gamme des produits proposés par l'ISI-WoS, ce sont les multiples usages du JCR, et de son algorithme JIF, qui ont favorisé leur quasi-exclusivité comme mesure de la qualité scientifique des revues et, par dérivation, des institutions et des chercheurs. Dans une telle configuration, l'évaluation des productions scientifiques se concentre presque exclusivement sur les articles publiés dans des revues centrales, c'est-à-dire celles indexées par l'ISI ayant un grand facteur d'impact. Elle dessine une autre économie du crédit scientifique dans laquelle la publication n'est plus uniquement destinée à produire des connaissances valides, mais à acquérir un important capital de citations. Et une telle perspective s'accompagne d'une conception particulière du cercle des pairs à même de juger de la qualité des travaux : relayant la définition confinée de l'expertise au cœur du processus d'évaluation dans les revues scientifiques, elle considère que les lecteurs sont nécessairement des auteurs citant d'autres textes (articles ou ouvrages) dans leurs propres articles.

4. Ce succès a fait de l'ISI une entreprise rentable qui a été rachetée par le groupe Thomson Scientific & Healthcare en 1992.

Toutefois, la position dominante de l'ISI-WoS a rapidement évolué au cours des années 2000 avec l'émergence d'autres producteurs de jeux de données, la percée d'algorithmes inédits et la diffusion d'outils bibliométriques alternatifs. Cette multiplication des entités se positionnant sur le marché de l'évaluation scientifique a déclenché le passage d'un monde monopolistique à une situation diversifiée. En d'autres termes, elle a engendré une toute nouvelle configuration.

2. La multiplication des producteurs de jeux de données

Le monopole acquis par l'ISI était le résultat d'un lent travail d'accumulation monétaire, logistique et commerciale. Pendant de nombreuses années, Garfield a continuellement fait la promotion de ses outils auprès de financeurs, de bibliothécaires, d'universités et de revues. Simultanément, Garfield (1979) appelait régulièrement à l'utilisation de tout jeu de données à des fins de production de connaissances. D'autres entrepreneurs ont tenté de faire exister leurs propres technologies de calcul, à l'instar de Francis Narin (Narin, 1976) qui a fondé Computer Horizons Inc. en 1968, afin de produire une bibliométrie explicitement évaluative, en partie financée par la National Science Foundation. C'est toutefois au début des années 2000 que deux véritables concurrents à l'ISI ont émergé : l'éditeur scientifique néerlandais Elsevier, et le géant du Web américain Google.

2.1. Scopus

Fondée en 1880 à Rotterdam, Elsevier est une maison d'édition scientifique qui a acquis à la fin du XX^e^ siècle, par le biais de rachats successifs, une position importante dans le domaine biomédical. En 2001, l'entreprise a créé Scirus, un moteur de recherche spécialisé dans les contenus scientifiques qui permettait, *via* un accès web, de faire des requêtes en texte intégral sur les articles publiés par Elsevier et par d'autres éditeurs commerciaux. Tout chercheur pouvait donc utiliser Scirus à des fins d'exploration bibliographique, sans abonnement payant. La devise de ce moteur de recherche, « *for scientific information only* », « seulement pour l'information scientifique », réincarnait l'objectif initial de l'ISI, mais dans une acception différente. Pour Garfield, la construction de jeux de données inédits était au service d'un fin repérage de l'information pertinente à partir d'un étalonnage de quelques revues. Avec Scirus, il suffisait de sélectionner, parmi les jeux de données existants, ceux contenant des informations scientifiques pour les agréger. Dans cette perspective, les sites web étaient traités comme des données et Scirus assurait éliminer ceux dont les contenus n'étaient pas scientifiques.

Elsevier a ensuite lancé Scopus en novembre 2004, un service payant intégrant les fonctions de Scirus et donnant également accès au texte intégral ainsi qu'aux articles citants et cités. Ce nouveau produit avait une ambition généraliste. À ce titre, il constituait la première véritable concurrence pour l'ISI-WoS : Elsevier et divers commentateurs soulignèrent les points communs et les différences entre les deux jeux de données, par exemple l'indexation

de l'ensemble des auteurs d'un même article et non du premier uniquement, ou la présence plus importante de productions dans d'autres langues que l'anglais.

Ce modèle semi-ouvert se fonde sur des échanges continus avec les utilisateurs, et perdure dans les développements successifs de Scopus, avec l'ajout, par les utilisateurs, de liens hypertextes vers le texte intégral de leurs publications. Ce sont également les utilisateurs qui peuvent suggérer, *via* une page web spécifique, des titres à inclure dans Scopus. Ces derniers sont sélectionnés par un comité, composé de vingt scientifiques et de dix bibliothécaires, sur la base de conditions d'inclusion comprenant l'existence de résumés en anglais, l'évaluation par les pairs et la parution régulière[5].

L'ISI, puis Thomson Reuters, ont régulièrement indiqué fonctionner selon une logique de quota, et ne retenir que 8 à 12 % des demandes d'inclusion sur la base du double argument de l'excellence de leur contenu et de la réduction de la masse d'informations autour de quelques revues centrales. À l'inverse, Elsevier focalise l'attention sur le très grand nombre des revues sélectionnées par son comité, présentant Scopus comme un dispositif ayant vocation d'exhaustivité. Loin de refuser des revues au contenu « exotique » comme l'ISI a pu le faire dans les années 1980 (Moravcsik, 1985), Elsevier ne cesse de mettre en avant l'extension de ses jeux de données vers des domaines peu couverts jusqu'alors (les sciences humaines et sociales, des supports dans d'autres langues que l'anglais, les brevets). L'enjeu n'est pas simplement cognitif, mais aussi commercial puisque Elsevier démarche les universités et les pouvoirs publics en affirmant que Scopus offre une meilleure couverture de l'activité scientifique et de son rayonnement et, par là même, permet une meilleure évaluation de cette activité.

2.2. Google Scholar

En novembre 2004, l'entreprise Google a également mis en place un jeu de données associé à un moteur de recherche spécialisé, Google Scholar. Comme Scirus, il est accessible *via* un site internet et permet de faire des requêtes sur les auteurs, les titres, les résumés. Une différence concerne cependant le critère « *for scientific information only* » qui n'est pas explicitement repris. Google Scholar ne sépare pas *a priori* les sites « académiques » des autres, mais opère une sélection à partir des motifs lexicaux de références bibliographiques :

> Google Scholar trie les articles de la même manière que les chercheurs, en tenant compte de l'intégralité du texte de l'article, de l'auteur, de la publication dans laquelle l'article est paru et du nombre de fois où celui-ci est cité dans d'autres ouvrages universitaires[6].

Cette extension est confirmée dans les développements ultérieurs : le contenu des ouvrages numérisés dans Google Books était accessible dans le jeu de données Scholar fin 2004, et Google a entrepris un travail de numérisation de fonds de revues scientifiques en 2007, en complément d'accords avec des éditeurs scientifiques commerciaux. Google a également proposé sur son site des procédures permettant aux bibliothécaires et aux dépôts d'archives

5. http://www.elsevier.com/online-tools/scopus/content-overview#content-policy-and-selection.
6. http://scholar.google.fr/intl/fr/scholar/about.html.

d'inclure leurs fonds dans les données Scholar. Dans la droite ligne du modèle économique de Google, combinant gratuité des usages et vente d'espaces publicitaires plus ou moins personnalisés, tout texte numérisé est potentiellement indexé dans le jeu de données Scholar. Les utilisateurs ont alors accès à des productions en texte intégral, archivées sur des sites personnels ou dans des dépôts institutionnels, quels que soient leur format informatique, leur langue ou leur genre (rapport, acte de colloque, chapitre d'ouvrage, tribune dans la presse…), que leur contenu soit ou non réservé à des abonnés. Enfin, le jeu de données Scholar est indexé de telle sorte qu'il permet le suivi des productions citantes et citées (Giles, 2005), ce qui en fait un autre concurrent à l'ISI-WoS, même en l'absence de commercialisation directe d'un « produit » Google Scholar.

Cet espace concurrentiel est très visible dans les travaux de bibliométrie et de scientométrie : à partir de 2005, des articles ont mis en scène les trois jeux de données, avec différents objectifs. Certains cherchent à vérifier leur « qualité », leur couverture et leur périmètre respectifs (Gardner et Eng, 2005 ; Burnham, 2006). D'autres opèrent une comparaison directe entre les trois jeux de données, sur une discipline ou un domaine de recherche, pour en souligner la cohérence ou les écarts générés en termes bibliométriques (Meho et Yang, 2006). Cette comparaison peut déboucher sur la construction de jeux de données combinés, ou sur leur hiérarchisation en fonction des usages (Falagas *et al.*, 2008 ; Bar-Ilan, 2010).

La production de jeux de données à destination des chercheurs s'est fortement diversifiée ces dix dernières années, Elsevier et Google n'étant que les deux acteurs les plus visibles et pérennes. Nous aurions pu aussi évoquer le cas de Microsoft, le producteur américain de logiciels, qui a proposé avec Live Search Academic un service proche de ceux de Google Scholar entre 2006 et 2008. Cette pluralité, inséparable du développement de l'Internet à haut débit, a trois caractéristiques : l'extension de la liste des productions scientifiques visibles, l'accessibilité accrue des produits de la recherche, et le détachement grandissant entre lieu de publication et contenu, *via* notamment des formes de *citation tracking*.

Or l'accès à ces jeux de données a également eu un effet en retour sur les algorithmes, qui peuvent être dorénavant aisément testés ou mis en œuvre à une large échelle. Parmi toutes ces propositions, un petit nombre connaît un succès important du fait de leur association à divers jeux de données et, finalement, de leur incorporation dans des outils bibliométriques standardisés. À ce titre, le succès fulgurant du *h-index* (l'« indice-*h* »), proposé par le physicien Hirsch (2005) pour subsumer la valeur d'un auteur sous un nombre entier, est exemplaire. Il a été inclus dans un outil bibliométrique dès octobre 2006, à l'aide d'un petit programme (Publish or Perish, PoP) élaboré par Harzing, une professeure de management, rendant son calcul opérationnel sur le jeu de données Goolgle Scholar.

Symétriquement, la lente trajectoire de l'Influence Weight, un algorithme développé par le bibliomètre Narin qui tient compte du prestige du lieu de citation et renforce ainsi l'importance des revues, est tout aussi parlante. Contrairement au JIF, cet algorithme est fondé sur un modèle récursif de l'influence où toutes les citations ne se valent pas. Longtemps ignoré en scientométrie, du fait de la puissance de calcul limitée des ordinateurs, l'Influence Weight a été redécouvert en webométrie lors du développement du PageRank, l'algorithme de classement de Google (Page *et al.*, 1999 ; Cardon, 2013). C'est seulement après de nombreux débats en scientométrie (Bollen *et al.*, 2006) que cet algorithme a été inclus dans deux

outils bibliométriques stabilisés, à la fois dans le *SCImago Journal Ranking* élaboré à partir de 2007 sur le jeu de données Scopus, et dans l'ISI-WoS, sous le nom d'Eigenfactor proposé par l'ISI-WoS à partir de 2010.

2.3. Une deuxième configuration

L'entrée en scène de jeux de données inédits, associée à l'apparition régulière d'algorithmes, renseignent sur les principales caractéristiques d'une deuxième configuration des infrastructures informationnelles dédiées à l'évaluation des productions scientifiques.

Cette configuration est tout d'abord marquée par une multiplication des acteurs en présence. Depuis les années 1970, l'ISI présentait une offre diversifiée sous couvert d'un unique référent. Dorénavant, plusieurs entreprises élaborent des jeux de données plus ou moins ouverts et accessibles, et divers acteurs produisent régulièrement des algorithmes ou des outils bibliométriques. À l'inverse de la position monopolistique tenue par l'ISI-WoS jusqu'au milieu des années 2000, le monde de la bibliométrie est désormais peuplé par des producteurs de jeux de données en concurrence et un oligopole d'algorithmes « stars » traversant l'ensemble des milieux scientifiques.

Cette deuxième configuration est ensuite caractérisée par un découplage notoire des technologies de calcul, qui prolifèrent sans être nécessairement encapsulées les unes dans les autres. Des producteurs de jeux de données se spécialisent exclusivement dans cette activité, à l'instar de Scopus qui rend visible, sur son site, la contribution indépendante de *SCImago,* en tant que concepteur de l'algorithme et producteur des résultats. Symétriquement, des chercheurs peuvent développer des outils bibliométriques sans produire d'algorithmes ni confectionner de jeux de données, comme le PoP développé par Harzing. Plus radicalement, un même algorithme donne des résultats différents selon le jeu de données sur lequel il est calculé : pour un unique JIF par revue, on a désormais plusieurs *influence weights* et différents *h-index* (Bar-Ilan, 2007). Ces transformations sont d'autant plus visibles que la majorité de ces nouvelles données bibliométriques sont publiques et reproductibles.

Enfin, cette deuxième configuration amorce une diversification du marché de l'évaluation quantitative des productions scientifiques. Avec l'arrivée de nouveaux producteurs de jeux de données et l'élaboration d'algorithmes inédits, les outils bibliométriques élargissent la gamme des entités visées : non seulement la définition de ce qui constitue une revue prestigieuse s'est transformée avec la redécouverte de l'*influence weight*, mais les revues ne constituent plus le seul et unique étalon de mesure. Des algorithmes portent aussi sur les auteurs ou les *working papers*. Pour autant, un usage bien particulier domine largement dans cette deuxième configuration : on scrute uniquement les citations que reçoivent des publications dans d'autres travaux scientifiques (revues, ouvrages, rapports, etc.). Autrement dit, le périmètre des pairs se limite aux auteurs de textes de ce type.

L'émergence récente d'outils alternatifs préfigure néanmoins une troisième configuration, marquée par une ouverture encore plus grande vis-à-vis des entités visées, des mesures mobilisées, et des types de participants pris en compte. Elle inaugure non seulement des formes inédites d'estimation des productions scientifiques, mais elle redéfinit simultanément le cercle des pairs jugés pertinents.

3. Infrastructures et métriques alternatives : prémices d'une troisième configuration

À partir des années 1990, le développement des archives ouvertes comme jeux de données et lieu d'implémentation des algorithmes a ouvert l'horizon des innovations bibliométriques. Leur généralisation dans les années 2000 a favorisé une diversification encore plus grande des technologies de calcul. En effet, plutôt que de simplement concurrencer le modèle de l'ISI-WoS comme producteur de données ou promoteur de nouveaux algorithmes de citation, plusieurs acteurs ont opéré deux transformations majeures : d'une part, la production et le stockage décentralisé de jeux de données ; d'autre part, l'adoption de mesures alternatives portant exclusivement sur les articles.

3.1. RePEc

La force de l'ISI-WoS était de proposer des jeux de données multidisciplinaires. Parallèlement, il existait des initiatives de partage de références, de bases de résumés, voire de textes complets selon une logique disciplinaire. Outre le cas d'arXiv, archive ouverte développée en 1991 à Los Alamos pour le dépôt de manuscrits non publiés (*preprints*), initialement en physique des hautes énergies (Gunnarsdóttir, 2005), on peut donner l'exemple moins connu des Chemical Abstract Services fondés par l'American Chemical Society en 1907, dont la collection de résumés est passée de 12 000 à 6 000 000 en un siècle, et dont les services n'ont cessé de se diversifier par l'inclusion d'une nomenclature des substances, des articles complets et de leurs citations, et d'une base de données de brevets (Baker *et al.*, 1980). En sciences humaines et sociales, PsycINFO, créé par l'American Psychological Association, a inclus en 2001 des outils portant sur les citations. Dans cette veine, le projet Research Papers in Economics (RePEc), initié en 1996, est marqué par une particularité : c'est le premier à adopter une architecture décentralisée et mondiale consacrée initialement aux *working papers*.

Conçu par un groupe d'économistes britanniques, RePEc s'est appuyé sur une première infrastructure de *working papers*, WoPEc, fondée en 1993 avant l'apparition du World Wide Web. Avec l'avènement de ce dernier, le projet a articulé deux éléments distincts. D'un côté, chaque institution, chaque laboratoire pouvait développer son propre serveur et rendre disponibles les textes stockés. De l'autre, les productions archivées n'étaient pas directement accessibles, mais interrogeables *via* des applications dédiées : outre WoPEc développé en Grande-Bretagne, citons IDEAS au Canada, RuPEc en Russie, pouvant accéder à l'ensemble des serveurs configurés pour RePEc. Le jeu de données lui-même contenait uniquement des *working papers*, avant de s'ouvrir progressivement aux articles publiés, à la suite de partenariats avec les principaux éditeurs scientifiques commerciaux[7]. Au 1er septembre 2014, le jeu de données comprenait plus de 592 000 *working papers* et plus de 1 016 000 articles publiés,

7. Elsevier est aujourd'hui le premier contributeur à RePEc.

ainsi que 18 000 ouvrages référencés, 90 % du matériel étant téléchargeable (gratuitement ou avec un abonnement chez les éditeurs concernés).

L'infrastructure informationnelle associe donc une décentralisation de la production et un stockage des données initiales (articles, références, résumés...), articulé à la construction d'un jeu de données unifié sur lequel de nombreux services sont bâtis (Karlsson et Krichel, 1999). Parmi ceux-ci, plusieurs sont des outils bibliométriques. Le projet LogEC, hébergé par la Stockholm School of Economics, enregistrait des statistiques d'usages d'IDEAS et proposait dès 2001 un classement des téléchargements (*top download*) ou un classement des vues de résumé (*top abstract views*) des textes inclus dans RePEc. Il mettait donc en œuvre des algorithmes issus des webometrics, dénombrant simplement l'accès à des pages web ou à des fichiers. Pour l'une des premières fois, la mesure de l'usage ne se faisait donc pas en analysant des productions scientifiques citantes, mais en comptant des visionneurs de résumés et des téléchargeurs de *working papers*.

Simultanément, un registre des pages personnelles, articulant les coordonnées d'économistes aux lieux de stockage de leurs *working papers,* a été développé. En 2003, il a pris un sens différent : identifier les auteurs des documents de RePEc afin de leur fournir des statistiques mensuelles sur les pages vues et les téléchargements. Enfin, en 2004, à partir d'algorithmes développés par des informaticiens sur l'archive ouverte CiteSeer, les promoteurs de RePEc ont construit une mesure des intercitations sur l'ensemble du jeu de données. Cela permet d'appliquer des algorithmes déjà largement partagés comme le JIF ou l'*influence weight* pour les revues, pour les collections d'ouvrages ou pour les collections de *working papers*. Les calculs s'effectuaient évidemment sur la base des références des documents inclus dans RePEc. On peut aussi obtenir le classement par auteurs (*top authors*) en appliquant 35 algorithmes distincts sur le jeu de données, allant du nombre de pages publiées (qui tient compte du nombre de coauteurs) au nombre total de citations, en passant par le *h-index*, le *top download* ou le *top abstract views*[8]. De même, des agrégations sur les plus de 13 000 institutions répertoriées sont disponibles. L'ensemble des divers classements est mis à jour et archivé une fois par mois.

Les outils bibliométriques de RePEc articulent donc un jeu de données très ouvert à une multiplicité d'algorithmes. Ces derniers peuvent provenir des institutions pionnières dans les deux configurations décrites précédemment ou être relativement spécifiques, comme le nombre d'auteurs RePEc citant un économiste donné. L'infrastructure RePEc ne hiérarchise cependant pas les critères les uns par rapport aux autres. Au contraire, sa présentation et son ergonomie assument un pluralisme des outils, chacun étant disponible au même titre que les autres. Au traitement indifférencié des mesures correspond une absence de présélection des données : à la manière de Google Scholar, tout document déposé sur un serveur alimente le jeu de données, même s'il n'est jamais publié dans une revue ou téléchargé par les utilisateurs de RePEc. De plus, IDEAS ou RuPEc étant totalement libres d'accès, les visionneurs ou téléchargeurs, tout en étant décomptés, peuvent être des usagers non académiques.

8. Voir par exemple : https://ideas.repec.org/top/.

3.2. Métriques alternatives (*Altmetrics*)

Outre les mesures de l'influence et du prestige, celles de l'attention ont pris de l'importance avec les réseaux sociaux tels Twitter et FaceBook. Au-delà des pages web professionnelles et des blogs, des réseaux sociaux spécifiquement dédiés aux chercheurs sont désormais disponibles (Academia.edu, ResarchGate.net, Mendeley.com, Peerevaluation.org). Ils permettent eux aussi d'effectuer divers comptages et d'en afficher les résultats : nombre de vues, de *likes*, de téléchargements, de *followers*, d'« amis académiques »… Ces possibilités inédites ont donné lieu à une prolifération des mesures de la production scientifique (Van Noorden, 2010), et, comme pour d'autres environnements en ligne (Napster, TripAdvisor, Amazon…), transforment les traces d'usages en éléments calculables. Elles ont ainsi ouvert la voie à une conception de l'évaluation qui s'oppose à la sélection de quelques outils supposés robustes et pertinents, au profit de l'agrégation d'une pluralité toujours plus importante d'indices. Une partie des promoteurs de cette conception se sont retrouvés dans le manifeste *Altmetrics* écrit en 2011, valorisant une gamme de mesures allant bien au-delà du seul JIF[9].

La création de *PLoS One* en 2006 est exemplaire de ce mouvement[10]. Publiée par l'organisation à but non lucratif Public Library of Science (*PLoS*), qui promeut la diffusion gratuite des productions scientifiques, cette revue à vocation généraliste est l'expression d'un nouveau genre. D'un côté, les articles sont publiés au fil de l'eau et accessibles sans abonnement pour les lecteurs, leur coût étant payés par les auteurs ; de l'autre, tous les articles considérés comme « techniquement justes » sont publiés. L'évaluation par les pairs est donc réduite ici à un contrôle de qualité, tandis que l'importance, la portée et l'originalité d'un article sont évaluées par ses usages ultérieurs. Pour ses promoteurs, *PLoS One* n'a donc pas à être jugée comme une entité unifiée, à la manière d'une revue classique, mais comme un dispositif facilitant la diffusion d'articles qui doivent être jugés individuellement.

Pour rendre opérationnelle cette conception, depuis mars 2009, chaque article est accompagné sur sa page web d'une palette d'indices, intitulée *article level metrics* (ALM), mise à jour en temps réel. Parmi les ALM, figurent le nombre de vues de la version HTML, le nombre de téléchargements du fichier PDF, le nombre de citations dans Scopus, CrossRef, WoS, PubMed Central, Google Scholar, et le nombre d'occurrences sur CiteUlike, FaceBook, Mendeley, Twitter et Wikipedia. Les lecteurs peuvent également noter de une à cinq étoiles la pertinence d'un article, sa fiabilité et son style. Rendues possibles par l'électronisation du support, toutes ces mesures sont considérées comme de bien meilleurs indicateurs de l'impact d'un article que les outils portant sur les revues.

> Dans un monde d'articles, les comptages de citations étaient la manière la plus simple de quantifier l'effet d'un article. Tracer les usages d'un article – et qui l'utilisait – était simplement impossible. Il était difficile de mesurer la rapidité avec laquelle une nouvelle théorie ou un nouveau concept prenait à l'intérieur d'une communauté scientifique. Les *article-level metrics* ouvrent la voie à des mesures portant à la fois sur l'effet immédiat et sur la socialisation d'un article. Ce sont des composantes essentielles de l'impact, qui n'avaient pas été prises en compte auparavant[11].

9. http://altmetrics.org/manifesto/.
10. Cette revue est devenue la plus prolifique du monde avec 23 406 articles publiés en 2012, et 31 500 en 2013.
11. http://www.sparc.arl.org/initiatives/article-level-metrics.

D'abord mises en œuvre sur les revues du groupe *PLoS*, les ALM ont été depuis adoptés par d'autres éditeurs, dont une vingtaine de revues du Nature Publishing Group et certaines du BMJ Publishing Group Ltd, ainsi que par le jeu de données Scopus de l'éditeur Elsevier. Ce passage de la quantification d'un seul usage (la citation d'une revue dans d'autres par exemple) à celle d'une multitude pose la question de leur commensurabilité et de leur hiérarchisation. Ces différentes données, à la fois publiques et disponibles sous la forme d'un fichier XML téléchargeable, permettent le traitement des ALM sous un double rapport. Agrégées, elles sont destinées à apprécier l'impact total (*total impact*) de chaque article, considéré comme la véritable mesure de sa valeur. Assemblées et retraitées, ces données produisent également des mesures d'évaluation ajustées aux objectifs de différents utilisateurs (chercheurs, institutions de recherche, agences de financements…).

Les limites traditionnelles de mesures de l'usage par d'autres productions sont ici largement franchies : tout lecteur ou « citeur » *via* des réseaux sociaux, même non académique, est pris en compte. Le cercle des pairs est donc à la fois composé de producteurs/citeurs classiques décomptés dans de multiples jeux de données (ISI-WoS, Scopus, Google Scholar…), de bloggeurs ou de journalistes relayant des informations (*via* Twitter, Facebook…), de simples lecteurs d'articles pas nécessairement professionnels du monde scientifique, et de téléchargeurs ou stockeurs de données. Autrement dit, la communauté pertinente s'étend bien au-delà des auteurs d'articles publiés dans des revues centrales ou influentes.

L'évaluation des productions ne se fonde donc plus sur un processus de certification préalable et sur le rang d'une revue. Elle résulte ici de l'agrégation de multiples jeux de données délivrant aussi bien des mesures d'impact (citations) que d'attention et d'audience (usages). La valeur d'un article est ainsi placée dans les mains d'une variété de lecteurs. Dans ce monde dominé par ce que les acteurs nomment le *post-publication peer review*, une autre économie du crédit scientifique, fondée sur la distribution de l'expertise entre divers publics, est également à l'œuvre : l'autopromotion est une pratique attendue, et il devient légitime de promouvoir ses propres travaux sur des blogs scientifiques et des réseaux sociaux. Cette naissance du « publiant marketeur » installe une vision politique de la science et de l'évaluation bien différente de celle instituée dans les deux précédentes configurations.

Cette troisième configuration émergente se caractérise donc par une multitude de jeux de données ouverts qui renvoient à des conceptions d'usagers fortement contrastées, non hiérarchisées entre elles. Du point de vue des algorithmes, une double tendance se dégage : d'une part, le recyclage et le raffinement des outils déjà présents dans les deux premières configurations et, d'autre part, le comptage brut qui est incorporé dans de nouveaux outils bibliométriques (nombre de tweets par exemple). Parmi ces derniers, le « *top download* », déclinable sur toute entité (auteurs, revues, institutions…) et sur tout jeu de données électroniques incluant des documents, est sans doute le plus partagé par les acteurs centraux de cette troisième configuration, mais également par d'autres plus traditionnels, à l'image du classement des 25 articles les plus « chauds » (*top 25 hottest articles*) présenté par Elsevier pour chaque domaine. Cette configuration est marquée par une triple rupture : l'étalon de mesure historique que sont les revues est concurrencé par un déplacement des mesures vers les articles, l'importance des usages se substitue à celle accordée aux productions, et le cercle des pairs s'élargit à d'autres catégories d'acteurs.

Conclusion

Trois configurations se dégagent de l'exploration des infrastructures proposée ici. Leur émergence successive dessine une tendance à la diversification des acteurs qui investissent le marché de l'évaluation scientifique, la prolifération des technologies de calcul, et l'élargissement de la gamme des entités prises en compte. L'apparition séquentielle de chaque configuration n'engendre cependant pas la disparition de la précédente. Au contraire, ces trois configurations coexistent : non seulement les jeux de données, les algorithmes et les outils bibliométriques sont simultanément disponibles, mais les mêmes entités sont actuellement prises en compte par différentes métriques. L'offre des technologies de calcul est aujourd'hui multiple et protéiforme, au point que des faussaires investissent ce marché florissant et cherchent à vendre une variété d'indices[12].

Dans une telle situation, les conventions permettant de s'accorder sur la qualité des entités évaluées (articles, revues, institutions, chercheurs…) sont loin d'être partagées. Par exemple, les promoteurs de la *Declaration on Research Assessment* critiquent les outils bibliométriques fondés sur les citations de revues, comme le JIF, au nom d'autres mesures valorisant l'attention et l'audience[13], centrales dans la troisième configuration. Ils insistent également sur la nécessité d'utiliser des jeux de données bibliométriques publics, afin que les mesures qui en sont tirées puissent être vérifiées par tout un chacun. Et, de leur côté, les scientomètres sont tiraillés entre les limites inhérentes aux outils de la première configuration qu'ils connaissent bien et l'explosion des jeux de données des deuxième et troisième configurations qu'ils jugent à la fois bénéfiques et difficilement contrôlables (Cronin et Sugimoto, 2014).

Depuis le milieu des années 1950, les infrastructures de données et les mesures de citation avaient contribué à organiser une économie de la rareté de l'espace des publications scientifiques. Le prestige et la qualité des revues étaient étalonnés à partir de celles dont le JIF est maximal. Cette organisation de l'évaluation se fonde sur une stricte séparation entre l'épreuve de certification pour publier dans telle revue et la mesure de la diffusion à travers les indices de citation. Tandis que le jugement des pairs se focalise sur le manuscrit lors de la première étape, il concerne la revue dans son ensemble au cours de l'enregistrement des citations. La diversification du marché de l'évaluation des productions scientifiques, associée au déplacement récent vers les métriques attachées aux articles, change radicalement la donne. Tenir compte de ces mutations, c'est non seulement prendre acte des transformations de la valeur des revues, mais aussi de celles du processus d'évaluation par les pairs lui-même (Pontille et Torny, 2015), dans lequel les traces fortement contrastées des usagers sont promues en instance de validation et de diffusion des productions scientifiques.

12. Voir par exemple : scholarlyoa.com/2014/02/13/more-questionable-scholarly-metrics-are-emerging.
13. http://am.ascb.org/dora/.

Références

Aiuti F., Baroni C., Cao A. et Fantoni A. (1991), « Academic promotion in Italy », *The Lancet*, vol. 338, nº 8778, p. 1337.

Annual Review of Phytopathology (1975), « Preface », vol. 13, nº 1.

Archambault É. et Larivière V. (2009), « History of the journal impact factor: Contingencies and consequences », *Scientometrics*, vol. 79, nº 3, p. 635-649.

Baker D.B., Horiszny J.W. et Metanomski W.V. (1980), « History of abstracting at Chemical Abstracts Service », *Journal of Chemical Information and Computer Sciences*, vol. 20, nº 4, p. 193-201.

Bar-Ilan J. (2007), « Which h-index? — A comparison of WoS, Scopus and Google Scholar », *Scientometrics*, vol. 74, nº 2, p. 257-271.

Bar-Ilan J. (2010), « Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar », *Scientometrics*, vol. 82, nº 3, p. 495-506.

Bollen J., Rodriquez M.A. et Van de Sompel H. (2006), « Journal status », *Scientometrics*, vol. 69, nº 3, p. 669-687.

Bowker G.C. et Star S.L. (1999), *Sorting things out. Classification and its consequences*, Cambridge (Mass.), MIT Press.

Bradford S.C. (1934), « Sources of Information on Specific Subjects », *Engineering: An Illustrated Weekly Journal*, vol. 137, nº 3550, p. 85-86.

Burnham J.F. (2006), « Scopus database: a review », *Biomedical Digital Libraries*, vol. 3, nº 1, p. 1-8.

Cardon D. (2013), « Dans l'esprit du PageRank. Une enquête sur l'algorithme de Google », *Réseaux*, nº 177, p. 63-95 (en ligne : https://www.cairn.info/revue-reseaux-2013-1-page-63.htm).

Cronin B. et Sugimoto C.R. (dir.) (2014), *Beyond bibliometrics: Harnessing multidimensional indicators at scholarly impact*, Cambridge (Mass.), MIT Press.

Fabbri L.M. (1987), « Rank injustice and academic promotion », *The Lancet,* vol. 330, nº 8563, p. 860.

Falagas M.E., Pitsouni E.I., Malietzis G.A. et Pappas G. (2008), « Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses », *FASEB journal*, vol. 22, nº 2, p. 338-342.

Gardner S. et Eng S. (2005), « Gaga over Google? Scholar in the Social Sciences », *Library Hi Tech News*, vol. 22, nº 8, p. 42-45.

Garfield E. et Sher I.H. (1963), « New factors in the evaluation of scientific literature through citation indexing », *American Documentation*, vol. 14, nº 3, p. 195-201.

Garfield E. (1979), « Is citation analysis a legitimate evaluation tool? », *Scientometrics*, vol. 1, nº 4, p. 359-375.

Giles J. (2005), « Science in the web age: start your engines », *Nature*, vol. 438, nº 7068, p. 554-555.

Gross P.L.K. et Gross E.M. (1927), « College libraries and chemical education », *Science*, vol. 66, nº 1713, p. 385-389.

Gunnarsdóttir K. (2005), « Scientific journal publications: on the role of electronic pre-print exchange in the distribution of scientific literature », *Social Studies of Science*, vol. 35, nº 4, p. 549-579.

Hicks D. (1999), « The difficulty of achieving full coverage of international social science literature and the bibliometric consequences », *Scientometrics*, vol. 44, n° 2, p. 193-215.

Hirsch J.E. (2005), « An index to quantify an individual's scientific research output », *Proceedings of the National Academy of Sciences*, vol. 102, n° 46, p. 16569-16572.

Karlsson S. et Krichel T. (1999), « RePEc and S-WoPEc: Internet access to electronic preprints in economics », *Electronic publishing' 99*, Ronneby.

Lavie P. (2009), « The race for the impact factor », *Journal of Sleep Research*, vol. 18, n° 3, p. 283-284.

Martyn J. et Gilchrist A. (1968), *An Evaluation of British Scientific Journals*, Aslib.

Meho L.I. et Yang K. (2006), « A new era in citation and bibliometric analyses: Web of Science, Scopus, and Google Scholar », http://arxiv.org/pdf/cs/0612132v1.pdf.

Moravcsik M.J. (1985), *Strengthening the coverage of the third world science*, National Science Foundation.

Narin F. (1976), *Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity*, Cherry Hill, New-Jersey, Computer Horizons.

Page L., Brin S., Motwani R et Winograd T. (1999), « The Pagerank citation ranking: bringing order to the web », Technical Report, Stanford InfoLab, p. 1-17.

Pontille D. et Torny D. (2010), « The controversial policies of journal ratings: Evaluating social sciences and humanities », *Research evaluation*, vol. 19, n° 5, p. 347-360.

Pontille D. et Torny D. (2013), « La manufacture de l'évaluation scientifique : algorithmes, jeux de données, outils bibliométriques », *Réseaux*, n° 177, p. 25-61 (en ligne : https://www.cairn.info/revue-reseaux-2013-1-page-23.htm).

Pontille D. et Torny D. (2015), « From manuscript evaluation to article valuation: The changing technologies of Journal Peer Review », *Human Studies*, vol. 38, n° 1, p. 57-79.

Raisig L.M. (1960), « Mathematical evaluation of the scientific serial », *Science*, vol. 131, n° 3411, p. 1417-1419.

Rossner M., Hill E. et Van Epps H. (2007), « Show me the data », *Journal of Cell Biology*, vol. 179, n° 6, 1091-1092.

Solla Price D.J. (de) (1963), *Little Science, Big Science*, New York, Columbia University Press.

Taubes G. (1993), « Measure for measure in science », *Science*, vol. 260, n° 5110, p. 884-886.

Van Noorden R. (2010), « Metrics: a profusion of measures », *Nature*, vol. 465, n° 7300, p. 864-866.

Weingart P. (2005), « Impact of bibliometrics upon the science system: Inadvertent consequences? », *Scientometrics*, vol. 62, n° 1, p. 117-131.

Wouters P. (1999), *The Citation Culture*, *Ph.D. thesis*, Amsterdam, Université d'Amsterdam (en ligne : http://dare.uva.nl/search?identifier=b101b769-100f-43e5-b8d2-cac6c11e5bbf).

Les facettes de l'Open Data : émergence, fondements et travail en coulisses

Jérôme Denis
Sociologue, professeur au Centre de sociologie de l'innovation (CSI-i3, Mines ParisTech / CNRS)

Samuel Goëta
Sociologue, cofondateur de la coopérative Datactist

> Les efforts se sont concentrés sur le travail politique et technique de mise en place de projets d'*open data*, mais pas assez sur l'analyse de ces mouvements discursifs et matériels et sur leurs conséquences. Il en résulte que nous manquons de travaux étudiant des projets d'*open data* en actes, la description des assemblages qui les entourent et leur donnent forme, et les manières désordonnées, contingentes et relationnelles selon lesquelles ils se déploient[1]. (Kitchin, 2014, p. 66)

Open data et *big data* sont deux termes qui sont entrés dans le vocabulaire des débats sur le numérique entre 2008 et 2009. Le site data.gov est lancé en mai 2009 par le président Barack Obama un an après que le magazine *Wired* se soit enthousiasmé, dans son édition de mai 2008, pour les promesses de l'exploitation des données massives à « l'âge du pétaoctet ». Bien qu'*open data* et *big data* soient apparus quasi simultanément et que le discours commun les confonde souvent, les deux termes désignent des pratiques différentes de diffusion et d'exploitation des données.

L'ouverture des données publiques (Open Data) caractérise les pratiques proactives de publication de données produites dans le cadre d'une mission de service public et ne contenant pas d'informations personnelles. Aujourd'hui, elle est devenue une priorité des gouvernements au point que les chefs d'État réunis lors du G8 de 2013 en Irlande du Nord ont adopté une charte sur l'Open Data qui déclare que les données publiques de leurs administrations doivent désormais être librement réutilisables par défaut. Si jusqu'ici les administrations devaient légalement partager ces informations lorsque le public les demandait, les politiques d'Open Data incitent désormais les agents à publier les données d'eux-mêmes. Gestion de la dépense publique, contrôle de l'environnement, efficacité des transports, transparence des industries extractives, régulation du marché des taxis : dans des situations très variées, l'Open Data est présenté comme une solution à une multitude de problèmes.

1. Les traductions, sauf mention contraire, sont des auteurs de ce chapitre [NdE].

Les données publiées par les gouvernements se prêtent rarement au qualificatif de *big data* ou « données volumineuses », « données massives » ; dans la plupart des cas, ce sont ce que Rob Kitchin appelle au contraire des *small data* (Kitchin, 2014). D'une part, du fait de leur volume : le plus gros fichier publié sur le portail *open data* français, data.gouv.fr, tient sur un disque DVD et la majorité des données n'y dépassent pas le mégaoctet. D'autre part, ces données font l'objet de mises à jour périodiques, annuelles ou mensuelles, exceptionnellement quotidiennes ou en temps réel. À l'inverse, les bases de données massives sont actualisées en permanence. Autre particularité, les données du Big Data ne sont quasiment jamais exploitables gratuitement et librement en intégralité alors que les données ouvertes (*open data*) peuvent être intégralement utilisées par tous. On peut toutefois citer les données ouvertes et massives des instituts météorologiques des États-Unis et de Finlande, mais les cas similaires sont exceptionnels. Malgré ces différences majeures, les données ouvertes sont une des sources des services du Big Data qui peuvent les croiser avec d'autres bases de données.

L'Open Data se distingue aussi du Big Data par les modalités de son émergence. Cette forme d'action publique répond à des aspirations variées : plus grande transparence de l'État, croissance économique de l'industrie de la donnée, transformation des administrations, participation du public ou encore liberté de l'information. Cet impératif de partage et de mise à disposition au public n'est par ailleurs pas neutre, il engendre une évolution des pratiques de production et de diffusion de données au sein même des administrations. Leur publication en vue d'une réutilisation par des tiers n'est jamais purement mécanique. Elle repose sur une série de transformations qui reconfigurent en partie le travail des agents administratifs.

Dans ce chapitre, nous proposons de revenir sur les circonstances de l'émergence des mouvements, des pratiques et des politiques aujourd'hui rassemblés sous le terme d'*open data*. Dans un premier temps, nous rappellerons le contexte historique des premières initiatives d'ouverture des données publiques. Cette approche nous permettra d'identifier les promesses et les aspirations qui ont été associées à cette notion aux multiples facettes. Nous nous appuierons ensuite sur une ethnographie de l'ouverture de données publiques dans des administrations locales et nationales en France[2] afin de comprendre le travail effectué dans les « coulisses » de l'Open Data. Nous décrirons les principales opérations qui permettent concrètement l'ouverture des données publiques.

1. Du droit d'accès au droit de réutilisation des données publiques

Il convient d'abord de revenir sur les origines de ce qui est progressivement devenu une véritable injonction à ouvrir les données publiques. Le mouvement que l'on désigne aujourd'hui sous le terme d'Open Data est ancré dans des concepts, des pratiques et des

2. Cette ethnographie rassemble plusieurs types de matériaux, afin d'appréhender les multiples facettes du travail d'ouverture des données. Des entretiens approfondis ont été menés dans des collectivités locales, des institutions et une entreprise, auprès de personnes en charge de la mise en œuvre du programme *open data*, de gestionnaires de données et de responsables informatiques. La démarche a été complétée par l'observation de comités de pilotage de deux projets *open data* ainsi que l'observation participante d'un programme en cours de préparation. Enfin, une série de documents internes et externes (plaquettes, articles de presse, billets de blogs et commentaires) a été analysée.

contraintes réglementaires qui anticipent largement son avènement. Déjà, en 1789, la Déclaration des droits de l'homme et du citoyen stipule dans son article 15 que « la Société a le droit de demander compte à tout Agent public de son administration » faisant de l'accès à l'information publique un des fondements de la démocratie naissante (Vismann, 2008 ; Kafka, 2012). Une racine plus proche des politiques d'Open Data réside dans la notion d'Open Government apparue à la suite de la Seconde Guerre mondiale aux États-Unis (Yu & Robinson, 2012). Portée en grande partie par des journalistes, l'« ouverture des gouvernements » désigne ici la révélation par l'État des secrets de son fonctionnement. L'émergence de ce terme, devenu rapidement un slogan, doit être comprise au regard de la critique de l'opacité de l'armée américaine lors de la Guerre froide et particulièrement pendant la Guerre du Vietnam (Yu & Robinson, 2012). Autour de la notion de « gouvernement ouvert », les journalistes tentaient alors d'obtenir un droit d'accès à l'information publique, un « droit de savoir », qui est devenu progressivement la norme dans la plupart des démocraties occidentales. En 1966, le *Freedom of Information Act* signé par le président Lyndon B. Johnson donne à chaque citoyen américain le droit d'exiger de l'administration les informations publiques qu'elle détient. En France, la loi CADA du nom de la Commission d'accès aux documents administratifs, établit en 1978 un droit similaire qui permet à chaque citoyen d'exiger d'une administration la publication des informations publiques. L'exercice de ce droit s'effectue par le dépôt d'une requête officielle auprès de correspondants en charge de leur traitement, la commission pouvant arbitrer en cas de litige. Il est important de préciser que les données concernées par ces dispositions sont des informations publiques, produites dans le cadre d'une mission de service public ; les données personnelles ou individuelles ne sont pas concernées.

Le terme *open data* a émergé dans un autre domaine. Il est apparu en 1995 dans un rapport de l'Académie des sciences américaine suggérant le partage libre des données environnementales collectées par satellite[3]. Dans des sciences comme la botanique, la génétique ou l'astronomie, les pratiques de partage de données sont devenues monnaie courante, associées au développement de grandes infrastructures assurant la diffusion libre des données dans les communautés scientifiques (Bowker *et al.*, 2010 ; Edwards *et al.*, 2011 ; Strasser, 2012). Cette politique vise à compenser le coût élevé de la collecte des données et à faciliter leur contrôle par les pairs.

Mais la diffusion du terme d'*open data* dans le vocabulaire des débats autour des politiques numériques ne s'installe qu'à l'issue de la rencontre de Sebastopol, tenue en 2007 en Californie, rassemblant des activistes du numérique comme Aaron Swartz, le fondateur de Reddit et militant de l'accès libre aux publications scientifiques, Lawrence Lessig, l'avocat créateur des licences Creative Commons, ou encore Tim O'Reilly, à l'origine de la notion de « Web 2.0 ». Cette rencontre vise à définir des principes essentiellement techniques de la diffusion de données, dans le but de faciliter leur réutilisation par des machines. Ces principes exigent la publication volontaire des données dès leur production, telles que collectées et sans modification ultérieure. Tandis que la notion

3. National Academy of Sciences (1995), « On the full and open exchange of scientific data », http://www.nap.edu/readingroom.php?book=exch&page=summary.html (dernière consultation le 22 mars 2017).

d'Open Government portait sur la publication de données sensibles concernant des secrets de l'État, la rencontre de Sebastopol impose des principes techniques au processus de diffusion sans en interroger le « contenu ».

Cette rencontre a exercé une grande influence sur l'équipe en charge du numérique du président Obama. Le jour de son entrée à la Maison-Blanche, il signe un mémorandum sur l'*open government* qui exige que chaque administration publie un plan d'action sur la transparence, la participation et la collaboration avec la société civile. Cette initiative a abouti à la publication d'une série de sites qui permettent aux citoyens de participer et d'accéder à l'information publique, le plus connu étant data.gov, premier portail de diffusion de données gouvernementales ouvertes. Il s'ensuit, par mimétisme, une prolifération de portails qui diffusent des données publiques, notamment au Royaume-Uni et en France où sont inaugurés data.gov.uk en 2009 et data.gouv.fr en 2011. Ces portails, qui regroupent des jeux de données extrêmement variés, s'inspirent des principes édictés à Sebastopol, dont ils stabilisent la définition par la pratique.

Les critères de définition d'une donnée ouverte, établis à Sebastopol et consolidés par des organismes comme l'Open Knowledge Foundation (OKFN) ou la Sunlight Foundation, peuvent aujourd'hui se résumer à quatre points principaux. Premièrement, une donnée est réputée ouverte si « chacun peut l'utiliser, la réutiliser et la redistribuer aux seules conditions de citer la source et/ou de partager à l'identique » selon la définition de l'OKFN[4]. Cette dernière condition impose que la donnée reste ouverte dès lors qu'elle est réutilisée publiquement, pour préserver son caractère de bien commun. Deuxièmement, d'un point de vue technique, une donnée ouverte doit être lisible par les machines (*machine-readable*) et exploitable pour un traitement automatisé. Cette condition invite à publier les données dans des formats ouverts, dont les spécifications n'appartiennent pas à un acteur, pour éviter une dépendance des utilisateurs à l'égard d'un logiciel particulier. Elle impose également de ne pas publier de fichiers au format PDF pour éviter une extraction complexe des données. Troisièmement, à la différence des informations diffusées dans le cadre du « droit de savoir », les données sont publiées volontairement par les administrations sur des portails qui facilitent leur utilisation par les usagers. Enfin, les données doivent être « brutes », c'est-à-dire diffusées sans retraitement, idéalement avec le plus haut niveau de détail, afin de réduire les asymétries d'information entre l'administration et le public.

Cette brève description de l'émergence de la notion d'*open data* relativise l'impression de nouveauté couramment associée à ce mouvement. Les politiques d'ouverture de données s'inscrivent dans la continuité de pratiques et de dispositions anciennes qui imposent la diffusion de l'information publique, complétées par des principes techniques qui favorisent le traitement automatisé des données. L'émergence de ces principes a été rendue possible par la convergence de plusieurs mouvements autour d'un concept porteur de multiples espoirs.

4. Open Knowledge Foundation, « The Open Definition », accessible en ligne : http://opendefinition.org/ (dernière consultation le 22 mars 2017).

2. Les multiples facettes de la demande d'ouverture des données

Les injonctions à l'ouverture des données se sont réclamées d'une multitude de mouvements et de principes. Concernant la demande d'ouverture des données publiques, cinq facettes peuvent être identifiées. Leur mise en évidence permet de dépasser l'apparente uniformité des acteurs qui se sont réunis autour du terme *open data*.

2.1. La transparence

Les politiques d'ouverture de données publiques sont couramment associées à la notion de transparence. Dans la continuité d'une série de mesures qui imposent à l'État de diffuser des informations concernant son fonctionnement, les politiques d'Open Data s'appuient sur l'idée que les institutions pourraient être transparentes et maîtrisables de l'extérieur. Dans la lignée de l'article 15 de la Déclaration des droits de l'homme et du citoyen, les constitutions démocratiques ont progressivement imposé aux gouvernements de rendre des comptes sur leur fonctionnement. Le droit de chaque citoyen à exiger des informations sur la conduite des affaires publiques, devenu la norme dans la plupart des démocraties, a suscité l'émergence du concept d'*accountability* (redevabilité) (Kafka, 2012). Depuis les années 1980, l'*accountability* désigne aussi la prolifération de dispositifs et de pratiques qui exigent que des comptes soient rendus et qui sont désormais tellement répandus que certains considèrent qu'ils constituent une véritable « culture de l'audit » (Porter, 1996 ; Strathern, 2000).

L'injonction à l'ouverture des données publiques s'inscrit directement dans la veine de ces dispositifs et dans le cadre de la mise en place d'objectifs et d'indicateurs généralisés. La publication de données ouvertes est ainsi souvent considérée comme une forme supérieure de transparence, réputée plus objective que les procédures de révélation qui l'ont précédée (Birchall, 2014). Dispensées d'un biais narratif ou interprétatif, les données « brutes » sont associées à une objectivité qui configure les politiques d'Open Data comme incarnant le renouveau voire le paroxysme de la transparence de l'État (Goëta, 2015). Par ailleurs, Evelyn Ruppert (2012) explique la dimension performative de ces politiques d'ouverture. L'utilisation des données ouvertes et leur circulation à travers des médiations sociotechniques met en œuvre (*enact*) la transparence publique, en faisant notamment émerger de véritables « publics des données » (*data publics*) qui sont institués par les dispositifs d'*open data* en inspecteurs du fonctionnement de l'État.

2.2. La libre circulation de l'information

La demande de données ouvertes s'inscrit également dans un mouvement plus large qui considère que la diffusion de l'information est impérative pour le maintien des sociétés démocratiques. L'influence de la pensée cybernétique est ici déterminante. En postulant que l'information est le « ciment de la société » indispensable à la perpétuation de la civilisation, la cybernétique défend en effet le modèle des sociétés « ouvertes » qui font reculer localement ce

que ces théoriciens appellent l'« entropie », la menace d'un chaos inéluctable qu'annoncent les principes physiques de la thermodynamique (Breton, 2004 ; Lafontaine, 2004 ; Triclot, 2008). L'ouverture des données publiques intervient par ailleurs à un moment où les données sont conçues comme une matière première (Ribes & Jackson, 2013) indispensable à la création de valeur ajoutée. Le schéma de la pyramide du savoir (Ackoff, 1989), admis couramment malgré son réductionnisme, place ainsi les données comme le fondement de toute forme d'information (Rowley, 2007 ; Frické, 2009 ; Floridi, 2011). Alors que les données sont désormais indispensables pour analyser un phénomène, leur diffusion est vue comme un vecteur de démocratisation de l'information et de l'expertise.

Les mouvements liés au logiciel libre ont aussi largement milité pour la liberté de circulation de l'information. Sébastien Broca, dans *Utopie du logiciel libre* (2013), revient sur les fondements intellectuels de ces mouvements et soutient que « la libre circulation de l'information est la valeur la plus communément associée au combat du Libre, que ce soit par les universitaires ayant étudié le sujet ou par les hackers eux-mêmes. » Des partisans du logiciel libre se sont en effet engagés pour que la connaissance et les œuvres bénéficient des mêmes libertés d'utilisation, de modification, de copie et de redistribution que le code informatique des logiciels qu'ils utilisent. Dans la continuité de luttes contre les entraves à la circulation de l'information, des « libristes » ont participé à la définition des principes de l'Open Data en défendant notamment l'usage de formats aux spécifications ouvertes et de licences dites *copyleft* qui empêchent légalement ce que la littérature autour des biens communs appelle l'« enclosure » des données. En France, « Regards citoyens », la principale association en faveur de l'ouverture des données publiques, utilise exclusivement des logiciels libres et comprend parmi ses membres fondateurs des militants de l'April, une association de défense du libre. Dans la revendication de la circulation de l'information, l'ouverture des données publiques apporte aussi une nouvelle exigence : celle de pouvoir réutiliser des données brutes et non des agrégats déjà traités par leurs producteurs.

2.3. La demande de données brutes

Les politiques d'ouverture des données publiques sont liées à l'entrée dans le langage courant de la notion de « données brutes ». Celle-ci est issue des sciences expérimentales et désigne les mesures telles qu'elles sont produites par les instruments. L'adjectif *brut* qualifie des données qui n'ont pas fait l'objet d'un traitement humain ou informatisé pour corriger ou « nettoyer » ces mesures (Walford, 2013). Dans *Science and Technology Studies*, Bowker (2008) et Gitelman (2013) ont dénoncé « un oxymore et une mauvaise idée » qui sous-entend que la production de l'information serait transparente et naturelle. Car, de fait, l'étude de la production de données scientifiques rend compte de l'impossibilité d'identifier un moment où la donnée aurait connu un état brut, sans traitement (Ribes et Jackson, 2013). Les « données » sont toujours « obtenues » (Latour, 1993, p. 188), elles sont le fruit de choix, de théories et d'interventions et ne sont jamais pré-interprétatives comme le laisse penser l'expression « données brutes ».

La revendication d'accéder aux données brutes va aussi de pair avec l'émergence des *data-driven sciences*, comme l'astronomie, la botanique ou la génétique, où le partage et la

réutilisation de grandes bases de données sont devenus des pratiques courantes (Strasser, 2012). Dans un contexte où les résultats scientifiques sont régulièrement contestés, publier les données avant leur traitement et leur agrégation devient une pratique courante pour assurer la transparence des procédures des chercheurs. En climatologie, où des controverses ont mis en doute la réalité du changement climatique, les chercheurs ont dû répondre aux demandes des climatosceptiques qui ont exercé leur droit légal d'accès à l'information (*freedom of information*) pour obtenir les données brutes (Edwards, 2010).

Concernant les données publiques des administrations, la revendication d'accès aux données brutes émerge en 2007 suite à un article de Rufus Pollock[5], fondateur du réseau international Open Knowledge qui milite pour l'ouverture du savoir. Dans un billet de blog intitulé « Give us the data raw and give us the data now », il critique les sites web publics qui se concentrent sur la production d'une belle interface pour présenter les données. Il explique que les interfaces deviennent rapidement obsolètes ; selon lui, les administrations devraient d'abord publier les données brutes pour que le public ou l'institution créent les interfaces qui vont répondre aux besoins des usagers. Cette exigence s'est répandue véritablement en 2009 après une intervention de Tim Berners-Lee, le fondateur du Web. Lors d'une conférence TED dont la vidéo dépasse le million de vues[6], il demande au public présent de crier « nous voulons des données brutes ! » Il y évoque l'existence de données « inaltérées », qu'il suffirait de réclamer pour qu'elles soient ouvertes à tous. Cette conférence a eu une grande influence sur l'émergence de l'injonction politique d'ouverture des données, envisagée comme un processus simple qui n'a pas à répondre à des critères de qualité, les données étant disponibles à l'état brut.

2.4. L'industrie de la donnée

Le quatrième aspect qui caractérise l'émergence des politiques d'*open data* a trait à l'avènement d'une industrie de l'information qui a considéré l'accès gratuit aux données publiques comme un « avantage compétitif » dans son développement (Kitchin, 2014). Les promesses économiques de l'ouverture des données sont liées à l'émergence d'une industrie de l'information. Cette industrie est particulièrement développée aux États-Unis où aucune information publique fédérale ne peut être protégée par le droit d'auteur depuis 1988 (Ronai, 1997). Dans les années 2000, la Commission européenne s'est inspirée du cadre réglementaire états-unien et a évalué la valeur économique de la diffusion gratuite ou à un coût marginal des données publiques à plusieurs dizaines de milliards d'euros par an dans l'Union. En 2003, elle a adopté la directive *Public Sector Information* (PSI) pour inciter les États membres à diffuser gratuitement leurs données. La Commission européenne multiplie les études sur le potentiel économique de la libération des données, évaluant jusqu'à 200 milliards d'euros par an la valeur de leur

5. Pollock R. (2007), « Give us the data raw, and give it to us now », *The Open Knowledge Foundation Blog*, http://blog.okfn.org/2007/11/07/give-us-the-data-raw-and-give-it-to-us-now/ (consulté pour la dernière fois le 27 février 2017).
6. Berners-Lee T. (2009), « The next Web », *TED*, http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html (consulté pour la dernière fois le 27 février 2017).

circulation optimale dans les pays de l'Union (Vickery, 2011). Ces promesses de croissance justifient les efforts financiers et humains (suppression des redevances, développement de portails, organisation d'événements) consentis par l'État pour ouvrir ses données. L'État planifie un retour sur investissement dans la taxation des revenus de la réutilisation dont il est prévu que le montant dépasserait largement le coût de l'ouverture.

Ces prévisions de croissance se sont transformées en revendication à la suite d'une campagne du quotidien britannique *The Guardian* en 2006 intitulée « Rendez-nous les joyaux de la couronne » qui exigeait la gratuité des données « payées par les impôts ». Parmi ces « joyaux », les données de l'Institut géographique britannique *Ordnance Survey* ont été perçues comme les plus prometteuses en termes de croissance. La tribune insistait sur le fait qu'aucun des services de cartographie numérique comme Google Maps n'était britannique du fait des fortes redevances exigées pour utiliser les données de l'*Ordnance Survey*. Le *Guardian* évoquait alors la différence avec la conception américaine :

> Aux États-Unis, on considère que les données collectées avec l'argent du contribuable devraient lui être restituées gratuitement. Et une étude détaillée montre que l'attitude fermée du Royaume-Uni envers ses données signifie que nous perdons des opportunités commerciales et même que nous freinons la recherche scientifique dans des domaines comme le changement climatique[7].

Cet extrait montre bien que la promotion d'une ouverture des données publiques au Royaume-Uni s'est formée autour de conceptions assez éloignées des autres facettes : nulle part il n'est fait référence aux notions de transparence ou de circulation de l'information.

2.5. La modernisation des administrations

Au-delà des promesses de croissance, les politiques d'*open data* reposent aussi sur leur capacité à provoquer des changements dans les procédures utilisées par l'administration et ainsi à accélérer sa « modernisation ». Les projets d'ouverture des données publiques sont soutenus par les services en charge de la transformation du service public. Ces derniers prévoient que la diffusion des données permettra aux agents de bénéficier des informations des autres services. La réutilisation des données produites dans d'autres divisions de l'organisation rendrait possible la fin du « travail en silos », expression désignant le cloisonnement des services et l'insuffisance des pratiques de partage d'informations. L'ouverture des données publiques offrirait aussi la possibilité de créer des services innovants en utilisant des données qui n'étaient jusqu'alors pas connues. Les administrations en charge de l'efficience du service public anticipent aussi que les producteurs vont devenir redevables auprès de leurs collègues de la qualité des données.

Cette modernisation concerne aussi le public lui-même. L'ouverture des données encouragerait la participation des citoyens à l'élaboration des politiques par la réduction des asymétries d'information. En disposant des données, idéalement dans leur intégralité

7. Charles et Cross (2006), « Give us back our crown jewels », *The Guardian*, 9 mars 2006, https://www.theguardian.com/technology/2006/mar/09/education.epublic (dernière consultation : 3 avril 2017).

et collectées selon les principes de Sebastopol, les politiques d'*open data* faciliteraient l'apparition d'une citoyenneté informée au fondement de politiques publiques d'inspiration libérale (Barry, 2001). La disponibilité des données publiques permettrait l'émergence de contre-pouvoirs qui pourraient avoir recours à celles-ci dans leurs activités militantes. D'autres modalités d'action collective ont vu le jour récemment, comme le « statactivisme » de certains acteurs qui se mobilisent pour que les données, les catégories et leurs indicateurs prennent en compte des situations jusqu'alors occultées ou ignorées (Bruno, Didier & Prévieux, 2014). Les services gouvernementaux en charge de la transformation de l'administration soutiennent les politiques d'*open data* pour favoriser une meilleure prise en compte des avis des citoyens. Plus largement, l'ouverture de l'information stratégique au public est perçue comme une source d'efficacité et d'innovation des organisations. Les politiques d'*open data* s'inscrivent dans la continuité des travaux des théoriciens des organisations qui encouragent une diffusion de l'information. En particulier, les travaux d'Henry Chesbrough (2006), à l'origine de la notion d'innovation ouverte, ont répandu l'idée qu'ouvrir l'information stratégique peut constituer une source d'innovation bien supérieure au régime classique du confinement des ressources en interne.

Ces bénéfices internes et externes positionnent les politiques d'*open data* comme moteur de la transformation du service public. La Banque mondiale encourage ainsi financièrement les pays en développement à ouvrir leurs données à la fois pour faire émerger une société civile informée, mais aussi pour « moderniser » le fonctionnement des administrations. Cette « modernisation » s'inscrit aussi dans le contexte de réduction des dépenses de l'État, où l'ouverture des données financières permettrait un contrôle de l'action publique qui dépasse les instances dédiées et l'étende à la société civile dans son ensemble. Des critiques ont associé l'ouverture des données publiques à la privatisation du service public (Bates, 2012). Une étude au Royaume-Uni a montré que les élus locaux ont réduit la question de l'ouverture des données à l'obligation de publication du détail des dépenses publiques de plus de 500 £ imposée par le gouvernement Cameron, oubliant les autres bénéfices attendus de la politique *open data* (Halonen, 2012).

Mais la progressive émergence d'une injonction à ouvrir des données dans les administrations publiques et la mise en œuvre de politiques d'*open data* dans de nombreux pays laissent paradoxalement dans l'ombre la question des données elles-mêmes et celle des conditions de leur « ouverture ». De quelles données parle-t-on exactement lorsqu'on défend la nécessité de libérer les données publiques ? Quelles sont ces ressources que certains dépeignent comme le pétrole du XXIe siècle, qu'il suffirait de mettre à la disposition de la population dans leur état « brut » pour qu'elle y trouve des moyens de produire de la transparence et de l'innovation ? Plus qu'un angle mort, les données constituent une évidence des politiques de l'*open data*. Insister pour que soit mis en œuvre un programme d'ouverture de données suppose que des données soient déjà présentes dans les administrations publiques, prêtes à être « libérées ». Or, dès que les programmes prennent chair dans les administrations, cette évidence ne tient plus. L'ouverture des données n'a rien de mécanique et la mise en œuvre concrète d'une politique d'*open data* passe par une série d'opérations qui témoignent de l'épaisseur sociotechnique des processus d'ouverture.

3. Le travail d'ouverture des données

Sans prétendre à l'exhaustivité, nous proposons d'exposer ici cinq étapes principales qui ponctuent cette ouverture progressive. Ces opérations montrent que les données, loin d'être des ressources toutes prêtes dont la diffusion pourrait s'effectuer à coût quasi nul, sont travaillées dans le processus même de leur ouverture. Elles sont d'autant plus délicates qu'elles sont portées par des collectifs de travail hétérogènes. Dans la plupart des cas étudiés, les chargés de projet sont rattachés à des services de communication et étaient au départ peu sensibilisés à la gestion des données administratives. Par ailleurs, au sein des services internes, ces politiques ajoutaient de nouvelles tâches et de nouvelles missions à des personnes qui ne travaillaient pas à la diffusion publique des données qu'elles produisaient ou qu'elles manipulaient. Les principaux acteurs de ces projets ne sont donc pas initialement des « petites mains » de l'information (Denis & Pontille, 2012) : leur activité déborde largement les seules production et maintenance de données et leur statut au sein des organisations concernées ne les vouaient pas à rester en coulisses d'un service destiné à des usagers extérieurs. En revanche, une partie des opérations qu'ils réalisent constituent bien un « travail invisible de l'information » (Denis, 2011), sur lequel reposent concrètement les politiques d'*open data* et leurs promesses de transparence, d'innovation, et de modernisation publique.

3.1. Identifier

Les moments qui initient les premiers pas d'un programme d'*open data* témoignent clairement du statut ambigu, et plus complexe qu'on ne l'imagine de prime abord, des données. Non seulement le périmètre des données potentiellement candidates à l'ouverture est débattu pour chaque projet concret, mais la connaissance même de leur existence, de leur nature et de leur emplacement est également problématique. Avant la question « comment allons-nous ouvrir tel ou tel jeu de données ? », de nombreux acteurs se confrontent à une interrogation plus générique : « de quelles données disposons-nous ? »

Les premiers pas d'un programme d'*open data* prennent ainsi la forme d'une exploration, une enquête menée par les personnes qui en ont la charge. Le terme d'*exploration* n'est pas anodin : cette enquête n'est pas un recensement. L'identification des données s'effectue progressivement, au fil d'échanges avec les services internes. Et le processus se nourrit lui-même, faisant émerger de nouvelles pistes au fil de son avancement. Selon les situations que nous avons étudiées, la sollicitation de ces services prend des formes variées : appel à initiatives internes, contact direct, premières idées de données pertinentes dessinées par l'équipe *open data* puis proposées aux services concernés, voire propositions faites par des associations de citoyens ou des développeurs qui ont eu l'occasion de les formuler.

L'inventaire, loin d'être mécanique, est un travail collectif : il engage des réunions, voire des négociations, qui articulent des enjeux hétérogènes, où se mêlent la concurrence entre les collectivités, l'intérêt imaginé du public pour tel ou tel type d'informations, la facilité ou au contraire les difficultés techniques liées à la diffusion des données ou à leur caractère plus ou moins « sensible ». Les données ne sont donc pas sélectionnées sur la base de critères

simples, définis à l'avance : leur exploration est progressive et collective. Elle passe par une co-construction qui montre que l'identification même des données qui vont être ouvertes participe pleinement du processus d'ouverture.

> On descend, on descend jusqu'au plus petit dénominateur commun pour qu'on puisse identifier vraiment toutes les données. Et ce qui est fou, c'est qu'à partir de ces trente rendez-vous, à chaque fois que je les rencontre, ils m'identifient cinq autres personnes qu'il faudrait que je vois donc, en gros, c'est un peu exponentiel. (Chef de projet *open data* d'une collectivité locale)

L'identification initie donc une opération générative (Law, 2009), un processus d'instauration, qui engendre un périmètre de données qui sont désignées non seulement comme « ouvertes » (ouvrables, dans un premier temps), mais aussi comme « données » tout court.

Cette première opération a également des conséquences organisationnelles. L'identification concerne autant des données que des services qui vont en devenir les référents. La mise en place de l'Open Data « travaille l'organisation », pour reprendre les termes de Cochoy, Garrel et de Terssac (1998), redistribuant certaines cartes, attribuant des rôles nouveaux et des responsabilités inédites.

3.2. Extraire

L'identification n'est qu'une première étape qui, dans la plupart des cas, ne suffit pas à « mettre la main » sur les données, notamment parce qu'une grande partie d'entre elles est littéralement enfouie dans des bases de données.

Le propre des bases de données relationnelles est d'offrir à leurs usagers des accès spécifiques, des « vues », répondant à leurs préoccupations professionnelles (Dagiral et Peerbaye, 2013). Ces « vues utilisateurs » évitent à tout un chacun de se préoccuper de la manière dont les données sont effectivement organisées dans les disques durs des machines (Castelle, 2013). Mais ce qui est pensé comme une source de confort et de facilité d'usages du point de vue des concepteurs des bases de données apparaît comme une contrainte forte pour les responsables d'un programme d'*open data*. Ouvrir les données suppose de passer outre les « vues utilisateurs » de chaque métier afin d'entrer en contact direct avec les données. Mais rares sont les bases qui disposent de fonctionnalités d'extraction automatique qui permettraient de récolter les données indépendamment de leur mise en forme logicielle :

> Il faut bien comprendre, c'est qu'au départ pour la plupart des systèmes des applications qu'on a chez nous qu'on a acheté, elles ne sont pas du tout conçues pour faire de l'*open data*. Donc, c'est compliqué. *On est obligé, nous, de développer des moulinettes, des tas de choses pour pouvoir sortir des données proprement*. (Gestionnaire de bases de données de transport)

L'extraction des données repose ainsi sur une autre forme d'exploration, qui passe par le développement d'outils *ad hoc* (des « moulinettes ») pour atteindre la « vue physique » des bases de données, outils d'autant plus complexes à élaborer qu'ils concernent des parcs informatiques hétérogènes, au sein desquels les logiciels ne sont que rarement harmonisés et où cohabitent des générations différentes d'outils.

C'est une deuxième étape de la progressive instauration des données qui se joue dans le travail d'extraction de données. Cette étape ne se limite évidemment pas à des aspects « purement » informatiques, mais elle dépend aussi des relations que les administrations entretiennent avec leurs prestataires techniques. Selon les termes des contrats, et la plus ou moins bonne volonté de ces prestataires, les vues physiques (les chemins d'accès vers les données « elles-mêmes ») sont plus ou moins masquées, et les bricolages pour y parvenir plus ou moins assimilables à des détournements, voire à des ruptures contractuelles. L'idée que les données publiques seraient des ressources dormantes qui ne demanderaient qu'à être libérées pour être exploitées, des « *commodities* » (Ribes *et al.*, 2013), est donc mise à mal non seulement par le coût que représente le travail d'extraction, mais aussi par l'ambiguïté de l'agencement sociotechnique dont les données sont dépendantes. Certains prestataires techniques qui vendent et entretiennent les systèmes d'information des institutions publiques sont propriétaires des chemins d'accès et des dispositifs de stockage de leurs bases de données : l'inaccessibilité des données est au cœur de leur modèle économique.

3.3. Nettoyer

Une fois les données identifiées et effectivement accessibles, elles passent par une étape cruciale : leur nettoyage. En sciences, ce type d'opérations est bien connu. Nombreux sont les travaux en *science and technology studies* qui ont insisté sur cette opération cruciale (Edwards, 2010 ; Walford, 2013). Si certaines tâches en sont proches, le nettoyage mis en œuvre dans les programmes d'*open data* ne répond toutefois pas aux mêmes préoccupations qu'en sciences : il n'est pas question ici de corriger les parasites générés par les instruments et le contexte de la mesure (en effaçant le bruit et les artefacts), mais plutôt de « mettre en qualité » des données en se confrontant à leurs usages précédents.

Cela passe par le repérage et la correction d'erreurs, par l'élimination de lignes vides, l'absence de données pouvant créer des incompréhensions, voire gripper le mécanisme des logiciels dédiés à les traiter. Il s'agit aussi pour les personnes chargées de cette étape d'harmoniser des identifiants, c'est-à-dire de travailler à l'articulation de jeux de données :

> Typiquement, sur les jeux de données des élections : entre les derniers fichiers des dernières élections, et puis les vieux trucs, les fichiers n'étaient pas présentés pareil. C'était des choses très bêtes mais il y avait des fois le titre de colonne qui était soit le nom du candidat soit le nom de son parti ou alors les deux, et j'ai essayé d'uniformiser tout ça pour que tous les fichiers se ressemblent et soient structurés pareil. (Chargé de projet *open data* dans une intercommunalité)

Quel est le sens de ces opérations de nettoyage ? Faut-il en conclure que les données avec lesquelles travaillent les administrations sont de mauvaise qualité, « sales », et que les programmes d'*open data* sont de bonnes occasions pour, enfin, les améliorer ? La question est un peu plus complexe. Comme l'ont montré Garfinkel et Bittner (1967) dans un article célèbre, la question de la qualité n'est pas une propriété intrinsèque des données, et l'on trouve de nombreuses « bonnes raisons organisationnelles » à l'existence de « mauvaises » données. C'est aussi ce que montre Lampland (2010) à propos des pratiques comptables qui

reposent en partie sur la manipulation de « *false numbers* » qui n'ont pas de conséquences particulières tant qu'ils ne sont pas saisis par des dispositifs d'audit qui en transforment le sens et la portée, les rendant problématiques. Les « mauvaises » données des uns peuvent ainsi être les « bonnes » données des autres. C'est le cas par exemple dans les transports, où les horaires utilisés dans certaines bases affichaient « 25 h 10 » pour « 01 h 10 » dans la nuit. De telles valeurs qui permettent de lever les ambiguïtés dans le cadre de la coordination du travail n'ont pas de sens en dehors de ce contexte.

Le nettoyage mis en œuvre dans le cadre des politiques d'*open data* ne témoigne donc pas d'une mauvaise qualité « intrinsèque » des données des administrations, mais plutôt du fait que l'ouverture constitue une épreuve pour les données qui, dans les mains d'autres usagers, pourraient être jugées inadéquates. En faisant migrer les données dans un cadre nouveau, les programmes d'*open data* rendent potentiellement centrales certaines de leurs dimensions qui étaient peu pertinentes dans leurs cadres d'usages initiaux. Des absences jamais remarquées deviennent des manquements, des approximations sans importance deviennent des erreurs, des redondances utiles deviennent des sources de problèmes.

3.4. « Brutifier »

Au nettoyage s'ajoute une opération plus curieuse, qui est en quelque sorte son prolongement, et qui témoigne de l'importance du travail sur les données en vue de leur ouverture. Nous l'avons vu, parmi les principes sur lesquels s'appuient les politiques d'*open data*, l'idée que les données ouvertes doivent être « brutes » est centrale. De bonnes données ouvertes sont des données qui ne sont pas retouchées. La première série d'opérations que nous venons de décrire met déjà à mal ce principe, puisqu'elle montre que, au contraire, l'ouverture des données engendre un certain nombre de manipulations essentielles à leur identification, à leur accessibilité et à leur qualité. Mais nous avons appris au fil de l'enquête que les données faisaient également l'objet d'une « brutification ». Certaines opérations visent en effet à les rendre brutes, comme les personnes qui en ont la charge nous l'ont expliqué, assumant ainsi pleinement ce que Geoffrey Bowker a qualifié d'« oxymore » à propos des données brutes en sciences (Bowker, 2000 ; Gitelman, 2013).

Cette « brutification » s'apparente à une forme de nettoyage mais ne vise pas à corriger des « erreurs ». Elle s'articule autour de deux enjeux étroitement liés : l'effacement des traces d'usages dans les données, et leur « délocalisation ». L'effacement des traces d'usage passe par exemple par le masquage de commentaires, ou celui de couleurs dans certaines cellules des tableaux. La délocalisation revient à faire disparaître les marques de l'ancrage des données dans des activités professionnelles spécifiques. Il s'agit par exemple de remplacer les nombreuses abréviations dont on sait qu'elles sont au cœur d'une forme d'efficacité langagière au travail (Fraenkel, 1994), ou encore de traduire des termes techniques dans un langage ordinaire :

> C'était un fichier Excel qu'ils avaient mis en forme selon ce dont ils avaient besoin. […] Donc, c'était vraiment leur fichier de travail. Or, nous, on ne voulait pas ça. Nous, on voulait des données plus brutes c'est-à-dire pas de commentaires, pas de tableaux, pas de mise en forme, juste vraiment les données au jour le jour, statistiques. Moi, je me suis occupée de ce travail-là, rebrutifier les données en fait. (Chargée de projet *open data* dans une intercommunalité)

D'un point de vue général, les opérations de « brutification » visent donc l'intelligibilité des données. Cette mise en intelligibilité passe évidemment par l'élaboration de métadonnées essentielles à tout projet de partage de données (Baker et Bowker, 2007 ; Edwards *et al.*, 2011) : dictionnaire, commentaires dans des documents à part, sont associés aux données pour que leurs usages soient facilités. Mais elle passe aussi par la transformation des jeux de données eux-mêmes, au sein desquels des termes vont être remplacés, des intitulés simplifiés, d'autres assemblés.

La « brutification » montre que le processus d'ouverture ne consiste pas à « libérer » des données qui resteraient primaires, mais à produire des données génériques, à vocation universelle. Cette production passe par une série de transformations par lesquelles des informations et des données « métier » deviennent des données ouvertes.

3.5. Formater

Enfin, il faut insister sur une dernière opération – le formatage – qui achève l'instauration des informations initialement identifiées en données ouvertes. Si l'ouverture des données publiques ne connaît pas le même degré de standardisation que le partage des données en sciences, il existe toutefois quelques standards et des formats de données qui sont en voie de consolidation et qui s'imposent dans les bonnes pratiques du secteur (Goëta et Davies, 2016). Parmi ceux-ci, le format CSV (*comma-separated values*) joue un rôle central aujourd'hui, notamment parce qu'il est considéré par la plupart des acteurs du domaine comme un format ouvert, compatible avec la plupart des logiciels tableurs, y compris du grand public. Nous avons ainsi pu entendre à l'occasion de la réunion de lancement du programme *open data* d'une organisation internationale le chargé de projet dire cette phrase qui souligne l'importance de la mise au format dans le processus d'instauration des données ouvertes : « pour moi du brut c'est du CSV ».

Or, la traduction d'un jeu de données en fichier CSV ne va pas de soi. Chaque exportation d'un format propriétaire issu des bases de données originales, ou d'un logiciel tableur, vers le format désiré réserve son lot de problèmes et donne lieu à des ajustements pour que les données initiales ne soient pas corrompues. Dans le meilleur des cas, des opérations en amont de l'exportation permettent de produire des fichiers qui seront compatibles et supporteront mieux le reformatage :

> — Quand tu passes [un fichier Excel] en CSV tout saute. Il y a des cellules fusionnées, du gras… Dans certains fichiers les gars ont mis de la couleur et la couleur a une signification, alors, que dans le CSV il n'y pas de couleur. Donc, tu es obligé de créer d'autres colonnes. […]
>
> — Et tu arrives à comprendre dans tous les cas, tout ce qui est… ?
>
> — Si on ne comprend pas on s'en réfère au producteur de données qui nous a envoyé le fichier. Mais sinon, tu vois, on reçoit aussi des fichiers avec des adresses, ce n'est pas géocodé. Du coup, on géocode. (Chargé de projet *open data* dans une région)

Par ailleurs, le formatage ne représente pas seulement une énième manipulation technique des données dans le processus qui mène à leur ouverture. Les standards engagent en effet toujours des enjeux politiques étroitement liés à leurs aspects les plus techniques (Lampland et

Star, 2008). Une partie des opérations précédentes décrites ici sont orientées vers cette mise au format des données, elles doivent en tout cas s'articuler avec celle-ci. La question des formats qui vont s'imposer et devenir des standards de l'*open data* est donc cruciale puisqu'ils vont non seulement définir, ou redéfinir en partie, la liste des acteurs institutionnels du domaine, ancrer les politiques d'*open data* dans des principes « de fait », traduits techniquement, mais également organiser le travail de l'information qui assure en coulisse l'existence même des données ouvertes (Goëta, 2014).

Conclusion

Tout au long de ce chapitre, nous avons montré que ce que l'on désigne aujourd'hui sous le terme d'Open Data regroupe des projets d'ouverture de données qui s'inscrivent dans un mouvement relativement récent et nourri de principes variés, inspirés notamment de la cybernétique, du mouvement du logiciel libre et du droit d'accès à l'information publique. Ces projets articulent par ailleurs des principes fort différents, au premier rang desquels la transparence politique, le soutien à l'innovation, et la modernisation des administrations. Au cœur des politiques d'Open Data se trouve toutefois un élément commun, dont l'existence semble évidente aux yeux de tous : les données. Les vocabulaires de la *libération*, de l'*ouverture* et de la *mise à disposition*, associés à l'importance nouvelle prise par les notions de données « brutes » ou « primaires », attribuent une place très particulière à ces entités informationnelles dont l'existence semble aller de soi, et dont la diffusion quasi mécanique est présentée comme un moteur de progrès.

En insistant sur certains aspects pratiques de la mise en œuvre de ces projets, nous avons montré, à rebours de cette vision désincarnée, que l'ouverture des données publiques reposait sur un travail complexe, sensible, par lequel des informations hétérogènes étaient progressivement instaurées en *open data*. Au fil d'une série de transformations techniques, linguistiques et politiques, s'opère ainsi la mue de données « métier » en données « ouvertes ». Nous avons également vu qu'au-delà du coût, généralement caché, que représente un tel travail, le processus d'ouverture a des répercussions organisationnelles qui dépassent les projets de modernisation habituellement envisagés.

L'ancrage historique, l'horizon des motifs qui animent les projets politiques et les activités concrètes qui nourrissent les projets donnent donc à voir la richesse, mais aussi la fragilité, des politiques d'Open Data, dont les sciences sociales ont tout intérêt à étudier la multiplicité plutôt que les grandes lignes communes. Il nous semble qu'une telle démarche serait également bénéfique à l'analyse des projets se réclamant des *big data*, dans un contexte où les débats contemporains tendent justement à se focaliser sur le traitement des données en aval, sur leur utilisation ou sur les enjeux de leur circulation. Au nom de quoi, par qui et dans quelles conditions ces données sont-elles produites ? Quelles transformations subissent-elles pour être associées les unes aux autres et devenir « massives » ? Autant de questions qui permettraient d'éclairer sous un nouvel angle l'assemblage sociotechnique des *big data*, dont tout le monde s'accorde aujourd'hui à souligner le caractère sensible sur les plans cognitif, politique et moral.

Références

Ackoff R. (1989), « From data to wisdom », *Journal of Applied System Analysis*, vol. 15, p. 3-9.

Baker K.S. et Bowker G.C. (2007), « Information ecology: Open system environment for data, memories, and knowing », *Journal of Intelligent Information Systems*, vol. 29, nº 1, p. 127-144.

Barry A. (2001), *Political machines. Governing a Technological Society*, Londres, The Athlone Press.

Bates J. (2012), « "This is what modern deregulation looks like": Co-optation and contestation in the shaping of the UK's open government data initiative », *Journal of Community Informatics*, vol. 8, nº 2.

Birchall C. (2014), « Radical transparency? », *Cultural Studies ↔ Critical Methodologies*, vol. 14 nº 1, p. 77-88.

Bowker G.C. (2000), « Biodiversity datadiversity », *Social Studies of Science*, vol. 30, nº 5, p. 643-683.

Bowker G.C. (2008), *Memory Practices in the Sciences*, Cambridge (Mass.), MIT Press.

Bowker G.C., Baker K., Millerand F. et Ribes D. (2010), « Toward information infrastructure studies: Ways of knowing in a networked environment », *in* Hunsinger J., Klastrup L. et Allen M. (dir.), *International Handbook of Internet Research*, Dordrecht, Springer, p. 97-117.

Breton P. (2004), *L'Utopie de la communication : le mythe du "village planétaire"*, Paris, La Découverte.

Broca S. (2013), *Utopie du logiciel libre. Du bricolage informatique à la réinvention sociale*, Neuvy-en-Champagne, Le Passager clandestin.

Bruno I., Didier E. et Prévieux J. (2014), *Statactivisme. Comment lutter avec des nombres*, Paris, Zones.

Castelle M. (2013), « Relational and non-relational models in the entextualization of bureaucracy », *Computational Culture*, nº 3, http://computationalculture.net/article/relational-and-non-relational-models-in-the-entextualization-of-bureaucracy.

Chesbrough H. (2006), *Open Business Models: How to thrive in the new innovation landscape*, Cambridge (Mass.), Harvard Business School Publishing.

Cochoy F., Garel J.-P. et de Terssac G. (1998), « Comment l'écrit travaille l'organisation : le cas des normes ISO 9000 », *Revue française de sociologie*, vol. XXXIX, nº 4, p. 673-699.

Dagiral É. et Peerbaye A. (2013), « Voir pour savoir. Concevoir et partager des "vues" à travers une base de données médicales », *Réseaux*, nº 178-179, p. 163-196 (en ligne : https://www.cairn.info/revue-reseaux-2013-2-page-163.htm).

Denis J. (2011), « Le travail de l'écrit en coulisses de la relation de service », *Activités*, vol. 8, nº 2, p. 32-52.

Denis J. et Pontille D. (2012), « Travailleurs de l'écrit, matières de l'information », *Revue d'anthropologie des connaissances*, vol. 6, nº 1, p. 1-20.

Edwards P. (2010), *A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming*, Cambridge (Mass.), MIT Press.

Edwards P., Mayernik M. S., Batcheller A.L., Bowker G.C. et Borgman C.L. (2011), « Science friction: Data, metadata, and collaboration », *Social Studies of Science*, vol. 41, nº 5, p. 667-690.

Floridi L. (2011), *The Philosophy of Information*, Oxford University Press.
Fraenkel B. (1994), « Le style abrégé des écrits de travail », *Cahiers du français contemporain*, nº 1, p. 177-194.
Frické M. (2009), « The knowledge pyramid: A critique of the DIKW », *Journal of Information Science*, vol. 35, nº 2, p. 131-142.
Garfinkel H. et Bittner E. (1967), « "Good" organizational reasons for "bad" clinic records » *in* Garfinkel H. (dir.), *Studies in ethnomethodology*, Englewood-cliffs, Prentice-Hall, p.186-207.
Gitelman L. (dir.) (2013), *« Raw Data » is an Oxymoron*, *Cambridge*, Cambridge (Mass.), MIT Press.
Goëta S. (2015), « L'Open Data : une forme ultime de transparence ? », *in* Catellani A. et Libaert T., *La communication transparente. L'impératif de transparence dans le discours des organisations*, Louvain-la-Neuve, Presses universitaires de Louvain, p. 49-66.
Goëta S. et Davies T. (2016), « The daily shaping of state transparency: Standards, machine-readability and the configuration of open government data policies », *Science and Technology Studies*, vol. 29, nº 4, p. 10-30.
Halonen A. (2012), *Being open about data. Analysis of the UK open data policies and applicability of open data*, The Finnish Institute in London, rapport disponible en ligne: http://finnishinstitute.cdn.coucouapp.com/en/articles/48-reports (dernière consultation le 27 mars 2017).
Kafka B. (2012), *The Demon of Writing. Powers and Failures of Paperwork*, Brooklyn, Zone Books.
Kitchin R. (2014), *The Data Revolution. Big Data, Open Data, Data Infrastructures & their Consequences*, Londres, Sage.
Lafontaine C. (2004), *L'Empire cybernétique. Des machines à penser à la pensée machine*, Paris, Seuil.
Lampland M. (2010), « False numbers as formalizing practices », *Social Studies of Science*, vol. 40, nº 3, p. 377-404.
Latour B. (1993), « Le pédofil de boavista, montage photo-philosophique », *Petites leçons de sociologie des sciences*, Paris, La Découverte, p. 171-225.
Latour B. (1993), *La Science en action*, Paris, La Découverte.
Lampland M. et Star S.L. (dir.) (2008), *Standards and their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*, Cornell, Cornell University Press.
Law J. (2009), « Seeing like a survey », *Cultural Sociology*, vol. 3, nº 2, p. 239-256.
Porter T. (1996), *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*, Princeton, Princeton University Press.
Ribes D. et Jackson S.J. (2013), « Data bite man: The work of sustaining a long-term study », *in* Gitelman L. (dir.), *« Raw Data » is an Oxymoron*, Cambridge (Mass.), MIT Press, p. 147-166.
Ronai M. (1997), « Données publiques : accès, diffusion, commercialisation », *Problèmes politiques et sociaux*, nº 68, p. 773-774.
Rowley J. (2007), « The wisdom hierarchy: Representations of the DIKW hierarchy », *Journal of Information Science*, vol. 33, nº 2, p. 163-180.

Ruppert E. (2012), « Doing the transparent state: Open government data as performance indicators », *in* Mugler J et Park S.-J. (dir.), *A World of Indicators: The Production of Knowledge and Justice in an Interconnected World*, Cambridge, Cambridge University Press, p. 51-78.

Strasser B.J. (2012), « Data-driven sciences: From wonder cabinets to electronic databases », *Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 43, n° 1, p. 85-87.

Strathern M. (dir.) (2000), *Audit culture. Anthropological studies in accountability, ethics and the academy*, Londres, Routledge.

Triclot M. (2008), *Le Moment cybernétique. La constitution de la notion d'information*, Seyssel, Champ Vallon.

Vickery G. (2011), *Review of Recent Studies on PSI Re-Use and Related Market*, Paris School of Economics, https://ec.europa.eu/digital-single-market/en/news/review-recent-studies-psi-reuse-and-related-market-developments (dernière consultation le 27 mars 2017).

Vismann C. (2008), *Files. Law and Media Technology*, Stanford, Stanford University Press.

Walford A. (2013), *Transforming Data: An Ethnography of Scientific Data from the Brazilian Amazon*, thèse de doctorat (PhD), Université IT de Copenhague.

Yu H. et Robinson D.G. (2012), « The new ambiguity of "open government" », *UCLA Law Review*, n° 178, p. 178-208.

III. Données numériques et outils de recherche en sciences sociales

Des données du Web pour faire de la sociologie… du Web ?

Jean-Samuel Beuscart

Sociologue, chercheur au sein du laboratoire SENSE (Sociology and Economics of Networks and Services) à Orange Labs et chercheur associé au Laboratoire interdisciplinaire Sciences Innovations Sociétés (Lisis, université Paris-Est)

Avec la croissance continue des usages d'Internet, une part grandissante de nos activités sociales se prolonge en ligne. Nos consommations et productions culturelles, notre sociabilité, nos curiosités et engagements politiques, une partie de nos achats, nos recherches d'information, nos demandes d'aide se développent de plus en plus dans les différents espaces du Web. Ces usages sont le plus souvent complémentaires des activités développées dans d'autres espaces sociaux, avec lesquelles elles sont fortement intriquées : l'information en ligne sur un produit est comparée avec celle donnée par un vendeur, et le produit est commandé sur Internet pour être livré en magasin ; on arpente les forums pour compléter l'information reçue lors d'une consultation médicale ; la musique écoutée en *streaming* convainc d'acheter un disque ou une place de concert ; le choix d'un restaurant est influencé par les notes que lui ont attribuées les internautes sur une application du téléphone ; parmi les amis avec qui nous interagissons sur Facebook, nous en croisons certains très régulièrement en face à face, tandis que nous n'en connaissons certains autres qu'à travers leur avatar sur le site. Pour une grande partie, ces activités sociales en ligne se déroulent dans des espaces publics ou semi-publics : les échanges, prises de paroles, votes, compteurs, y sont le plus souvent visibles de tous, à tout le moins d'un grand nombre d'abonnés ou « amis ».

Ces activités sociales en ligne produisent un très grand nombre de traces : liens « d'amitiés » de toutes sortes, *likes*, conversations, expressions d'avis, votes, notes, compteurs, achats, pages de profils renseignées par les utilisateurs, goûts déclarés, historiques d'achats, d'écoutes, de visionnages, etc. Du point de vue du sociologue, ces traces sont autant de données potentiellement « disponibles » très prometteuses. Elles sont issues des activités réelles des acteurs sociaux, produites en situation ordinaire, donc *a priori* exemptes à la fois des biais caractéristiques des enquêtes déclaratives et de ceux des situations expérimentales. Pour le sociologue de la culture, par exemple, l'accès aux compteurs de visionnages des vidéos vues sur le Web, ou à l'historique des consommations d'un utilisateur sur un service de *streaming*, permet d'avoir une information sur ce qui est vraiment vu et écouté, plutôt que sur les goûts déclarés (souvent plus légitimes). Les données issues du Web sont donc potentiellement mobilisables pour décrire des activités sociales, et contribuer à l'approfondissement des questionnements sociologiques.

Néanmoins, du fait de l'intrication complexe entre comportements hors ligne et en ligne, le traitement de ces données ne va pas de soi. L'observation des activités du Web procède comme une coupe dans l'entrelacement des activités, en ne retenant que les actions du Web qui ont laissé une trace dans les données. Or tous les acteurs sociaux n'investissent pas avec la même énergie les espaces en ligne dans chacune de leurs activités ; et il est souvent difficile de

relier les actions sur Internet aux actions hors du Web. Les données issues du Web constituent donc à la fois une opportunité inédite pour la sociologie, par la richesse des interactions en situation « naturelle » qu'elles décrivent, et un défi méthodologique, dans la mesure où elles sont presque toujours porteuses d'un biais de représentation non contrôlé. C'est à l'examen de cette question que se consacre ce chapitre : dans quelle mesure, et à quelles conditions, peut-on produire des énoncés sociologiques à partir des données issues du Web ?

La question est d'autant plus prégnante que les sociologues ont été devancés dans ce domaine par les chercheurs en sciences informatiques, plus familiers de ces données et de leur manipulation ; depuis une dizaine d'années, ils produisent des constats sur les mondes sociaux en ligne, qu'ils étendent parfois à l'ensemble des mondes sociaux. Ces travaux sont extrêmement riches et innovants, parfois conduits dans un esprit d'ouverture interdisciplinaire, mais ils sont souvent ambigus – ou excessifs – quant au statut de leurs énoncés sur la société. Une recension des différentes difficultés méthodologiques et ambiguïtés épistémologiques de ces recherches peut être utile, au moment où ces méthodes se diffusent dans les sciences humaines et sociales.

Pour avancer sur cette question, je m'appuierai d'une part sur les retours d'expérience de ma participation à des travaux sociologiques faisant intervenir des données du Web, au sein du laboratoire SENSE d'Orange Labs et dans le cadre de projets coopératifs (ANR PANIC, ALGOPOL). Au cours de ces travaux sur des objets variés (les pratiques culturelles amateurs, la consommation télévisuelle, le marché de la restauration, la recommandation culturelle, la conversation en ligne, etc.), nous avons souvent « recueilli » sur le Web des données qui semblaient pertinentes pour notre objet, avant de découvrir au fur et à mesure de l'analyse ce qu'elles permettaient – et ne permettaient pas – de formuler comme énoncé de sciences sociales[1]. D'autre part, je m'appuierai, dans un échantillonnage forcément imparfait et subjectif, sur les travaux en sciences humaines et en informatique réalisés au cours des dix dernières années à partir des données du Web. Concernant les recherches en informatique, je mobiliserai celles qui s'inscrivent le plus, explicitement ou non, dans une ambition de description du monde social, et de dialogue avec les sciences humaines et sociales, psychologie et sociologie en particulier. Je m'appuie tout particulièrement sur les éditions annuelles de la conférence ICWSM (*International Conference on Weblogs and Social Media*, créée en 2007), et sur les travaux les plus sociologiques de la conférence WWW, organisée annuellement depuis 1994 en partenariat avec le World Wide Web Consortium (W3C) pour réfléchir aux évolutions et aux impacts du réseau. Cela laisse de côté un très grand nombre de travaux, mais fournit un matériau suffisamment riche pour commencer la réflexion.

Ce chapitre s'organise en deux temps. La première partie montre dans quelle mesure les données du Web permettent de faire des constats très fins sur les espaces sociaux en ligne, en s'efforçant de souligner à la fois l'inventivité méthodologique des travaux, la finesse des descriptions des logiques sociales en ligne rendues possibles par les données, et les zones d'ombre qui leur échappent, qui ne peuvent être compensées que par la complémentarité des approches méthodologiques. La seconde partie s'attache ensuite à discuter dans quelle

1. Ces recherches collectives doivent beaucoup aux compétences de Thomas Couronné et Thomas Beauvisage, qui ont développé les outils d'extraction des données mobilisées.

mesure les constats produits à partir des données du Web peuvent prétendre à une validité sociologique générale, au-delà de la description des espaces sociaux en ligne. Elle identifie ainsi trois principales postures de recherche, qui sont autant de réponses à cette question : la mesure des « effets » du Web, le Web comme dispositif d'enquête, le Web comme reflet homothétique de la société.

1. Des constats précis mais incomplets sur les usages du Web

Par « données issues du Web », nous entendons les données construites à partir des traces des activités sociales en ligne des internautes, que celles-ci soient fournies par les sites web qui organisent ces activités, ou qu'elles soient construites par le chercheur à partir des informations visibles sur le Web. Après avoir rappelé les limites inhérentes au mode de production de ces données, nous soulignerons la variété et la richesse des constats qu'elles permettent de produire, dans le cadre général d'une sociologie des nouveaux médias. Le travail à partir des seules données du Web pêche néanmoins par manque d'épaisseur sociale des personnes, qui ne sont saisies qu'à travers des informations sociodémographiques et biographiques très parcellaires, et gagne à être articulé à d'autres méthodes.

1.1. « Extraire » les données : limites méthodologiques intrinsèques

Malgré le vocabulaire utilisé, qui laisse supposer que les données sont disponibles, « extraites » ou encore « aspirées », leur recueil est en fait toujours un travail de construction, qui engage des choix, des omissions, des conventions et des compromis. Comme le montrent Samuel Goëta et Jérôme Denis dans cet ouvrage, la « donnée » porte bien mal son nom, parce qu'elle n'est jamais donnée ; et la « donnée brute », censément la moins travaillée, directement extraite du Web, doit en fait toujours être « brutifiée », travaillée pour être détachée de ses inscriptions originelles[2]. Les méthodes de recueil de traces du Web n'échappent pas à ce travail de construction des données (Beauvisage, 2013).

Dans un premier cas, les données sont construites indépendamment des propriétaires des sites web concernés, au moyen d'un programme informatique *ad hoc* qui parcourt les sites, en extrait les données pertinentes, et les intègre à une base de données. Le programme automatique (*web-scraper* ou *crawler*) définit d'une part une heuristique de navigation, la façon dont il va passer d'une page à l'autre pour en lire les informations, pour s'efforcer d'avoir une vision la plus complète possible du site ; d'autre part, il repère les informations publiques jugées pertinentes – le plus souvent à partir de leur positionnement dans la page HTML – et les copie dans une base de données. La capacité du programme à « récupérer » des informations dépend de plusieurs caractéristiques du site. La structure du site, tout d'abord, rend plus ou moins aisée la navigation exhaustive (si tant est qu'elle est envisageable, ce qui dépend de la

2. Voir *infra*, chapitre 6.

taille du site et des moyens dont dispose le chercheur) : observer tous les membres ou tous les objets d'un site est beaucoup plus facile s'il en existe un annuaire ou une liste exhaustive, à partir desquels on peut simplement suivre les liens vers les objets ou les personnes. Dans le cas contraire, le programme peut par exemple suivre les liens sociaux entre les personnes, avec le risque d'oublier les comptes peu connectés avec les autres ; ou faire des recherches (par ville, par thématique, etc.), avec encore une fois des risques de « trous » et une impossibilité de prétendre à l'exhaustivité. Le second élément qui détermine la qualité de la base de données est bien sûr la nature des données visibles sur le site. Celui-ci ne montre souvent qu'une partie des informations renseignées par les utilisateurs, tout comme il ne rend visible qu'une partie des traces de leurs actions, celle qui est susceptible d'être pertinente pour la navigation et l'activité sociale des utilisateurs. Par exemple, la plupart des sites du Web social rendent publics le nombre de contacts et de gratifications (*likes*, favoris, *views*, commentaires, etc.). En revanche, la date de ces gratifications (le *timestamp*) n'est pas toujours renseignée, ce qui place sur le même plan, dans les données ainsi construites, les succès d'un jour et ceux construits de longue haleine. En outre, la richesse de la base de données dépendra de la politique du site lui-même quant à la navigation des robots : le fichier robot.txt peut interdire certains types de navigation, et l'administrateur du site peut exclure les programmes à une navigation trop intensive ou systématique. Enfin, dans le cas, par exemple, de constitution d'une base de données longitudinale, au moyen d'un programme relevant des compteurs à intervalles réguliers pour étudier la progression des informations, la base est tributaire des évolutions du design du site, le moindre aménagement entraînant l'invalidité du programme et de nouveaux « trous » dans les données (Beuscart et Beauvisage, 2012).

Dans un second cas de figure, les données sont construites par les administrateurs du site eux-mêmes, et non reconstruites à partir d'une navigation systématique. Les données sont alors susceptibles d'être plus riches et les traces de constituer un meilleur reflet de l'activité sociale en ligne, car elles peuvent inclure des informations pertinentes qui ne sont pas rendues visibles pour les internautes. Cette richesse est cependant limitée par plusieurs facteurs. Tout d'abord, les données « disponibles » dépendent à la fois des besoins, des moyens et de la maturité de l'entreprise gestionnaire du site. Les entreprises n'enregistrent pas toutes les traces de l'activité de leurs utilisateurs ; certaines traces ne sont pas « loguées », c'est-à-dire enregistrées de façon durable par le site. Les besoins des entreprises ne sont pas ceux des chercheurs, et les informations conservées ne sont pas nécessairement celles qui sont jugées pertinentes par le chercheur ; par exemple, l'administrateur d'un site peut ne pas avoir jugé pertinent d'enregistrer la date exacte de chaque gratification, ou le chemin d'accès emprunté par l'utilisateur pour parvenir à la visualisation de tel contenu. Ensuite, les entreprises peuvent être réticentes à fournir certaines données jugées stratégiques, ou potentiellement nuisibles à leur image ou à leur développement dans un contexte fortement concurrentiel. C'est notamment le cas des informations sur la fréquence de connexion des utilisateurs, et plus généralement des indicateurs d'intensité d'utilisation des services, qui montrent souvent qu'une majorité des utilisateurs d'un service n'en ont qu'un usage très occasionnel. Enfin, bien entendu, l'extraction et la construction de ces données par les entreprises, leur anonymisation, leur « nettoyage », représente un temps de travail non négligeable ; la décision d'allouer cette force de travail à la constitution d'une base de

données pour le chercheur dépend donc aussi de l'intérêt que l'entreprise perçoit dans le travail de recherche réalisé à partir de ces données.

Une configuration intermédiaire est celle de la constitution d'une base de données par le chercheur à partir des API (*application program interfaces*, interfaces de programmation) mises en place par les sites. Dans ce cas, la constitution de la base de données est à la fois plus facile, et plus contrainte par les données rendues disponibles ou non par le site, et par les conditions de leur extraction. Par exemple, un chercheur intéressé par la production et la consommation de vidéos en ligne peut bénéficier des API de YouTube, et recueillir les adresses IP, le nombre de vues, de *likes*, de commentaires, les mots-clés, etc., des vidéos liées à une recherche ; mais chaque requête est limitée à 1000 résultats, ce qui rend difficile toute prétention à l'exhaustivité, aussi étroit que soit le sujet traité. Ainsi, comme le notent Boyd et Crawford (2012), l'apparence et la revendication d'objectivité et d'exhaustivité qui accompagnent très souvent les analyses des larges données du Web doivent donc toujours êtres nuancées.

1.2. Des constats macroscopiques précis sur les usages du Web

Malgré leurs incomplétudes, les données issues du Web permettent de formuler des constats précis et solides sur les comportements sociaux en ligne, et constituent une assise d'une valeur inestimable pour la sociologie des usages des nouveaux médias.

Tout d'abord, ces données, notamment celles issues des grands sites du Web 2.0, permettent de dessiner les contours et les reliefs des publics participatifs en ligne. La démocratisation et la massification des usages d'expression et de participation sur le Web, à travers ses sites emblématiques (Blogger, YouTube, MySpace, Flickr, Twitter, Tumblr, Instagram…) et leurs concurrents de niche, s'est accompagnée d'un discours célébrant la participation et l'expression de tous les internautes dans l'espace public. Certaines versions de ces discours étaient porteuses des excès et généralisations hâtives caractéristiques des enthousiasmes technologiques. L'étude des données des différents sites ont permis une description nuancée de ces usages, en en restituant à la fois l'étendue, la diversité, les inégalités et les hiérarchies.

Ainsi, les premières études sur les blogs (Herring *et al.*, 2005 ; Adamic *et al.*, 2005), YouTube (Cha *et al.*, 2007), Flickr (Mislove, *et al.*, 2007), tout en confirmant la forte appétence des internautes pour ces sites (les aspirations de Flickr en 2006, deux ans après sa création, contenaient 4,5 millions de comptes et 150 millions de photos), ont mis en évidence la très forte hétérogénéité des engagements des participants. Mathématiquement, la distribution de la participation est souvent décrite par une loi de puissance : beaucoup d'utilisateurs n'ont qu'une très faible activité, tandis qu'un très petit nombre d'utilisateurs participe de façon très intense. Seuls 5 % des contributeurs de Wikipédia ont réalisé 10 *edits* ou plus (Levrel, 2006) ; la plupart des contributeurs à YouTube ont posté moins de 2 vidéos, tandis que 0,5 % en ont posté plus de 1000 (Cha *et al.*, 2007) ; sur Flickr, 20 % des utilisateurs fournissent 80 % des photos (Beuscart *et al.*, 2009). L'analyse de ces données met également en relief la diversité des usages qui sont faits des différentes fonctionnalités proposées par le site. Par exemple, parmi les usagers intensifs de Flickr, certains mobilisent très peu les fonctionnalités sociales

du site et l'utilisent comme un espace de stockage de photos, tandis que d'autres utilisent intensivement les outils de conversation, sans poster aucune photo ; seule une petite proportion des usagers correspond à la figure de l'internaute 2.0, postant ses œuvres et participant aux discussions sur ses créations et celle des autres.

Les travaux sur les données ont également fourni une image macroscopique des régularités de la sociabilité foisonnante qui se développe sur ces sites. Les outils de l'analyse des réseaux permettent de caractériser les échanges sociaux sur ces sites en termes de réciprocité, de distance moyenne entre deux individus, de degré entrant et sortant (nombre de contacts), de connectivité, etc. Mislove *et al.* (2007) ont ainsi mené une comparaison et une modélisation systématique des indicateurs de description de réseaux sociaux de différentes natures (Flickr, Orkut, LiveJournal, YouTube) ; ils mesurent la distance moyenne du chemin entre des individus pris au hasard (qu'ils évaluent entre 4 et 6 selon les sites), le degré de réciprocité des liens, la distribution des degrés, etc. Les travaux pionniers sur les réseaux de mails puis sur les blogs mettent en évidence que la structure des liens sociaux en ligne ressemble aux réseaux sociaux hors ligne, tout en en accentuant les caractéristiques. Les réseaux en ligne s'organisent ainsi en cliques plus ou moins denses, interconnectées entre elles, certains nœuds jouant le rôle de passeurs entre les univers ; la représentation graphique typique montre alors des groupes (*clusters*) de nœuds interconnectés, souvent dotés de caractéristiques proches, eux-mêmes plus ou moins proches et connectés à des groupes voisins. Kumar *et al.* (2006) ont ainsi représenté Flickr sous la forme d'une « composante géante » réunissant la majorité (2/3) des utilisateurs connectés entre eux, entourée d'une pluralité de « satellites » constituée de petites communautés denses mais séparées du reste du groupe, et d'individus isolés. Dans un article classique, Adamic et Glance (2005) dessinent la blogosphère politique américaine durant l'élection américaine de 2004 comme deux masses denses se faisant face : les blogs républicains sont fortement interconnectés entre eux, tout comme le sont les blogs démocrates. La communication entre les deux ensembles est assurée par des blogs « apolitiques », moins nombreux, qui font office de passeurs entre les deux univers. De même, les musiciens présents sur MySpace ont pu être représentés en fonction des liens de citations explicites qu'ils entretiennent entre eux : le graphe qui en résulte, coloré en fonction des genres musicaux déclarés, fait apparaître des zones de couleur très nettes, montrant le regroupement des artistes en cliques plus ou moins fermées représentant leur scène musicale (Beuscart et Couronné, 2009).

Par rapport aux réseaux sociaux *offline*, la distribution du nombre de contacts est cependant beaucoup plus inégale : des internautes peuvent cumuler plusieurs milliers de contacts (ou citations, ou amitiés, etc.), tandis que d'autres en reçoivent très peu (Barabasi, Ravasz et Vicsek, 2001). Ici encore, les premiers travaux sur les blogs (Herring *et al.*, 2005) ont présenté un constat largement repris et confirmé par la suite : les univers participatifs en ligne sont fortement hiérarchisés et structurés. Les blogs de faible renommée citent ainsi à la fois d'autres blogs peu connus (leurs semblables) et des blogs beaucoup plus lus ; en revanche, ces blogs reconnus abondamment cités par les autres ne se citent qu'entre eux, renforçant ainsi les différentiels existants de notoriété. Cette tendance a pu être constatée dans de nombreux autres espaces du Web, par exemple chez les musiciens de MySpace, dont les 10 % ayant le plus d'audience reçoivent plus de la moitié des liens, tout en n'émettant quasiment que des liens vers cette même élite de musiciens connus (Beuscart et Couronné, 2009).

Ce constat d'inégalité de connectivité des internautes participant aux différents sites peut être élargi en un constat d'inégalité d'attention portée à leurs productions. Dans la foulée des travaux sur la structure du Web (Adamic et Huberman, 2000), les recherches mobilisant les compteurs de marques explicites d'attention (vues, *likes*, commentaires, favoris, etc.) observent toujours la distribution très inégale (en loi de puissance) de l'attention des internautes, qu'il s'agisse de blogs, de vidéos YouTube (Cha *et al.*, 2007), de photos (Beuscart *et al.*, 2009), de musiciens (Stoica *et al.*, 2010), de tweets (Cha *et al.*, 2010), d'émissions de télévision en *replay* (Beuscart et Beauvisage, 2012), etc. Une minorité de productions concentre l'essentiel de l'attention, tandis que la plupart des prises de paroles ou créations ne reçoivent qu'une attention faible. En outre, dans le cas des produits culturels, les biens les plus populaires en ligne sont ceux produits par les industries culturelles qui bénéficient d'importants budgets de promotion et d'une notoriété médiatique globale importante (Bastard *et al.*, 2012).

1.3. Une compréhension des trajectoires en ligne des individus et des contenus

Les données issues du Web permettent donc d'appréhender l'activité sociale des grands univers du Web, notamment de souligner la distribution très inégale de la participation comme de l'attention, et de proposer des représentations des composantes de ces univers. Au-delà de ce travail macroscopique, la recherche permet également une analyse fine des trajectoires des personnes et des textes sur le Web, à travers des questionnements qui font directement écho aux problématiques de la sociologie des médias : comment se construit la réputation ? Comment se diffuse une œuvre, une idée ? La notoriété est-elle durable ? Comment une opinion, une thèse, l'emporte-t-elle sur ses concurrentes ? Comment les thématiques s'imposent-elles dans l'espace public, pourquoi y restent-elles ? Ces travaux issus en grande majorité des sciences informatiques citent d'ailleurs fréquemment des classiques de la sociologie des médias tels que le *Personnal Influence* de Katz et Lazarsfeld (1955).

Il est illusoire de prétendre rendre ici la finesse et la richesse des très nombreux travaux qui s'attachent à ces questions à partir des données du Web. Nous en donnerons simplement quelques aperçus. Un premier apport est la compréhension des logiques de construction de la grandeur des personnes sur le Web, au-delà du constat de sa distribution. Cha *et al.* (2010) montrent ainsi que, sur Twitter, la réputation est (aussi) le produit d'une spécialisation et d'une expertise prolongée dans un domaine : tout comme les leaders d'opinion de Katz et Lazarsfeld, les influenceurs sur Twitter ne le sont que sur un domaine précis, où leur expertise s'est construite progressivement. Sur ce point, une littérature importante, ancrée notamment en sciences du marketing, s'est attachée à mesurer l'influence des personnes dans la circulation en ligne d'un bien ou d'une idée, et à nuancer l'idée que certaines personnes très connectées seraient « hyperinfluentes » sur le Web (Leskovec, Adamic et Huberman, 2006 ; Watts et Dodds, 2007 ; Godes et Mayzlin, 2009 ; voir aussi Beauvisage *et al.* [2011] pour une synthèse en français). Dans un autre registre, Cardon *et al.* (2011) montrent, à partir de données longitudinales sur les liens de citation des blogs de cuisine, qu'il existe « deux chemins de la gloire » pour les blogueurs et blogueuses : soit au travers d'une reconnaissance au sein de la communauté, soit par l'importation d'une notoriété acquise au dehors, auprès

des médias traditionnels notamment ; les deux trajectoires sont en grande partie orthogonales, et les exemples de conversion d'une réputation « interne » en réputation « externe » sont statistiquement rares.

Symétriquement, de nombreux travaux s'efforcent de retracer et d'expliquer la circulation des entités (vidéos, images, idées, expressions, rumeurs, publicités, etc.) sur le Web. Dans un article devenu classique, Leskovec *et al.* (2009) analysent la production de l'information durant la campagne présidentielle américaine de 2008, combinant données issues des sites de la presse en ligne et des blogs. À partir de l'identification de groupes de mots (qu'ils nomment *memes*) issus des discours des candidats, ils décrivent l'espace public en ligne américain comme une succession de pics d'attention autour de certains sujets. Dans cet espace, les médias traditionnels restent prescripteurs et sont « suivis » par les éditeurs en ligne. Plusieurs travaux se sont ensuite attachés à distinguer des formes distinctes de ces focalisations de l'attention, tel Lehmann *et al.* (2012) sur Twitter. Du côté de l'explication du succès, Cha *et al.* (2008, 2009) distinguent, dans le succès d'une photo sur Flickr, ce qui ressort vraiment de la recommandation entre pairs (la viralité au sens strict) et ce qui est imputable à d'autres formes d'exposition. Friggeri et ses collègues (2014), de leur côté, se sont intéressés à plusieurs types de cascades informationnelles, qu'elles portent sur des offres promotionnelles que les consommateurs sont incités à relayer, ou sur des rumeurs qui sont spontanément propagées sur Facebook. Dans ce dernier cas, ils observent que certaines catégories de rumeurs sont plus propices à la propagation sur Facebook (celles liées à la politique, à la médecine, à la nourriture, au crime) et que la viralité dont elles bénéficient est bien plus importante que celle observée sur d'autres types de contenus.

D'autres travaux enfin permettent de se situer au niveau des trajectoires individuelles, et de deviner à partir des données longitudinales les ressorts des comportements sociaux en ligne. Huberman *et al.* (2009) expliquent ainsi, à partir d'un large échantillon de créateurs de vidéos YouTube, la probabilité de persévérer dans cette activité par le succès des créations passées. Dans le même esprit, Prieur *et al.* (2008) montrent que le meilleur prédicteur du succès sur Flickr est le nombre de commentaires donnés : pour recueillir de l'attention, il faut en distribuer beaucoup. Liu *et al.* (2014), à partir d'une base de données de 37 milliards de tweets, décrivent l'émergence et l'adoption de conventions d'écriture et de partage sur Twitter, telles que la pratique du retweet ou l'usage de certains *hashtags*. Michael et Otterbacher (2014), reprenant un débat sur la dépendance de sentier dans les notes et avis en ligne suggérant que les premières évaluations d'un produit vont influencer les suivantes, montrent que, dans les termes employés et les critères mobilisés, les évaluateurs profanes ont tendance également à être influencés par les avis précédents.

Les données issues du Web permettent donc d'approcher finement les mécanismes de circulation des personnes et des textes en ligne, ainsi que ceux de la construction de leur grandeur. Les discussions autour de la « viralité » conduisent à préciser les modes de circulation des textes, les formes de grandeur des personnes, et le rôle de ces dernières dans la circulation des premiers. La granularité permet souvent d'approcher les trajectoires individuelles pour observer les formes de persévérance, d'abandon, d'adoption de conventions, de conformisme, etc.

1.4. Un manque d'épaisseur sociale

Les données issues du Web permettent une sociologie fine des usages des nouveaux médias, en fournissant à la fois des images macroscopiques indispensables à l'appréhension de ces phénomènes de masse, et des analyses fines et robustes des trajectoires des textes et des personnes. Néanmoins, par rapport aux exigences d'une compréhension sociologique, les données manquent souvent d'épaisseur sociale.

Tout d'abord, issues des traces laissées par les internautes dans leur activité, enregistrées et restituées par le site, les données du Web ont tendance à déformer la représentation qui est faite de l'activité en ligne. Les traces d'activité, logiquement, surreprésentent les plus actifs. Sur les sites participatifs et sociaux, les traces de conversation et d'appréciation explicites sont enregistrées, mais les traces de navigation silencieuse sont très rares. On peut connaître le nombre de *retweets* et de favoris d'un *tweet* ou d'un utilisateur, pas le nombre de fois où ses *tweets* ont été vus. D'un utilisateur de Flickr ou d'Instagram, on connaîtra les commentaires ou favoris qu'il a distribués, pas le nombre de photos qu'il a vues ; s'il a une navigation intensive mais inexpressive du site, il apparaîtra comme « inactif » selon les critères utilisés. D'une vidéo YouTube, on connaît le nombre de vues, mais pas le nombre d'internautes ayant visité la page ; etc. La représentation issue des données a toujours tendance à minorer les spectateurs discrets.

Plus encore, les données issues des traces manquent souvent d'informations quant aux inscriptions sociales et économiques des participants. C'est d'ailleurs principalement sous cet angle que les recherches sur les données du Web ont pu être critiquées jusqu'à présent. Julie Denouël et Fabien Granjon (2011) regrettent ainsi la pauvreté sociologique des informations mobilisées, et la forte abstraction sociologique des internautes décrits dans nombre de ces recherches. Il arrive qu'on connaisse leur âge et leur genre, avec une marge d'incertitude dans la mesure où il s'agit d'informations déclaratives. De même, il est souvent possible, avec une certaine marge d'erreur toujours, de différencier les amateurs des professionnels sur les sites relatifs aux pratiques culturelles ; ce fut par exemple le cas sur MySpace où, au prix d'un certain « nettoyage », les informations sur l'inscription professionnelle des musiciens (« major » / « indépendant » / « non-signé ») étaient utilisables et pertinentes pour l'analyse des comportements en ligne et du succès (Caverlee et Webb, 2008 ; Beuscart et Couronné, 2009). En revanche, il est extrêmement rare de disposer d'indicateurs du niveau d'éducation ou de revenus, de l'occupation professionnelle, ou de la classe sociale de ces participants en ligne, alors qu'il est incontestable que ces dimensions ont une influence sur la propension à s'exprimer en ligne et sur la façon de le faire. Danah Boyd et Kate Crawford (2012) prolongent cette critique en discutant la représentativité problématique des données. En soutenant que« les données les plus massives ne sont pas toujours les meilleures » (« Big Data are not always better data »), les auteures s'opposent aux chercheurs trop enthousiastes estimant que les grandes données du Web, en permettant d'appréhender la totalité des utilisateurs, rendent obsolète la réflexion sur la représentativité des données, pourtant commune à toutes les méthodologies des sciences sociales. Prenant l'exemple de Twitter, elles rappellent que, si large soit la base de données, il est toujours difficile de savoir exactement de qui on parle. Tout d'abord, les simples spectateurs (*lurkers*) sont considérés comme inactifs, et leurs

traces invisibles ; et la base de données comme les API sont soumises à des pannes et à des surcharges qui détruisent des informations. Surtout, les auteures notent que les utilisateurs de Twitter ne sont pas du tout représentatifs de la population, sans qu'il soit possible de corriger ou même de mesurer ce biais, et de savoir quels sont les groupes surreprésentés dans l'expression sur le site, etc. Une autre critique formulée par Boyd et Crawford vise le caractère désincarné et simpliste de certaines interprétations des indicateurs, alors qu'ils sont issus de traces d'activités dont le sens dépend du contexte. Par exemple, dans l'étude des liens sociaux en ligne, un lien d'amitié sur Facebook n'a pas le même sens qu'un contact sur Flickr, ou qu'un *follower* sur Instagram ou sur Twitter : dans le premier cas, les liens numériques recouvrent en partie les liens hors ligne ; dans le second, il s'agit essentiellement de liens en ligne et dissymétriques ; dans le troisième, la situation est très variable selon les utilisateurs. L'interprétation de la distribution de ces liens et de la dynamique de leur création ne saurait donc être homogène d'un site à l'autre. Le risque est sinon, pour reprendre la critique sévère de Denouël et Granjon, de

> […] confondre […] quelques-unes des traces des usages avec la vérité sociale des (non-) pratiques qui leur correspondent et qui restent indéductibles des seuls indicateurs à partir desquels ils travaillent (*hits* de page, nombre d'amis, de commentaires, etc.). […] La sophistication des moyens mis en œuvre, aussi impressionnante soit-elle, ne saurait cacher une certaine indigence de l'analyse sociologique (Denouël et Granjon, 2011, p. 36).

Dans le même ouvrage, Josiane Jouët (2011) note le danger d'une « réification des liens électroniques qui s'affranchit de l'appartenance à d'autres mondes sociaux » et appelle à un renouvellement de la critique autrefois portée par Charles Wright Mills des apories de l'empirisme dans les études des médias et de leur faible épaisseur sociologique.

Chez les sociologues ayant recours à ces données pour comprendre les activités en ligne, il existe plusieurs stratégies pour redonner une consistance sociologique à ces données, ou du moins pour observer les biais dont elles sont porteuses.

D'une part, il est possible de mesurer, à partir d'enquêtes statistiques plus classiques, des biais tels que le biais de participation, et plus généralement de construire une toile de fond à l'aune de laquelle interpréter les données du Web. S'appuyant sur l'enquête régulière de l'Oxford Internet Institute, Grant Blank s'attache ainsi à repérer les facteurs sociodémographiques qui prédisposent à la participation en ligne ; si l'âge et le niveau de diplôme jouent un rôle important, il souligne que c'est avant tout le degré d'expérience d'Internet (en ancienneté et en intensité) qui explique les comportements de contribution (Blank et Reisdorf, 2012 ; Blank, 2013). Plus généralement, de nombreuses études par questionnaire, notamment les travaux pionniers de l'université du Michigan, ou celles du Berkman Center, ont été menées pour essayer de ramener les usages observés en ligne à des ancrages sociaux tangibles en termes d'âge, d'éducation, de position sociale, de genre, etc.

D'autre part, les analyses des données du Web gagnent à être complétées par des approches qualitatives permettant de restituer le contexte et le sens social donnés aux activités en ligne. Les entretiens approfondis et les observations permettent de resituer les différents sens, intentions, espoirs qui guident l'attribution d'un *like* ou d'un favori, qui justifient l'énergie et le soin mis dans le *post* d'une photo ou d'un texte. Dans nos

travaux sur les musiciens ou sur les photographes, les entretiens avec les contributeurs des sites ont permis de dessiner la gamme des logiques sociales de la présence artistique amateur en ligne, renforçant ainsi les interprétations des données en termes de sociabilité et de notoriété (Beuscart, 2008 ; Crepel, 2011) ; ce sont d'ailleurs avant tout les entretiens qui ont permis la formulation d'une typologie des passages de l'amateur au professionnel, les données servant alors de toile de fond (Beuscart et Crepel, 2014). De même, les travaux sur les notes et avis de consommateurs en ligne combinent l'analyse statistique des données des sites – indispensable pour comprendre les formes de différenciation à l'œuvre dans ce dispositif d'évaluation – et des entretiens avec les acteurs (sites, professionnels, clients) nécessaires à la découverte des contextes de mobilisation de cette évaluation (Mellet *et al.*, 2014).

Cette combinaison des approches reste cependant une solution de second rang ; les approches statistiques et ethnographiques permettent de contextualiser les données du Web et donc d'orienter leur interprétation, mais pas de les qualifier directement. On sait que les plus diplômés sont plus susceptibles de tenir un blog, mais pas pour autant quel est le niveau de diplôme associé à tel ou tel type de prise de parole. Les protocoles de recherche permettant la qualification sociodémographique des avatars s'exprimant en ligne, et l'interview d'un sous-échantillon d'entre eux, sont rares et difficiles à mettre en place, et ne peuvent l'être que pour des volumes de données relativement réduits.

2. Trois postures méthodologiques pour faire des constats sur la société

Nous nous sommes concentrés jusqu'à présent sur les usages des données du Web qui visent à faire une sociologie du Web. Dans ces travaux, les données extraites permettent une compréhension des activités sociales en ligne, en mesurant les différentes formes de contribution, en soulignant les immenses variations d'engagement et de visibilité entre les usagers, en identifiant des typologies d'usage, des trajectoires d'engagement, constats idéalement complétés par d'autres matériaux permettant de contextualiser les interprétations.

Nous nous intéressons maintenant aux recherches s'appuyant sur les données issues du Web pour produire des constats généraux sur la société, plus ou moins explicitement inscrits dans les traditions des sciences sociales. Les traces d'activité en ligne concernent en effet des activités sociales telles que les pratiques culturelles, la sociabilité, le jugement de goût, la consommation, les sentiments politiques, etc., et peuvent être analysées non pas seulement comme les indices d'un comportement en ligne, mais aussi comme des indicateurs de comportements sociaux plus généraux. Les *Web data* viennent alors – timidement – enrichir l'arsenal méthodologique des sciences sociales, sans rester cantonnées aux spécialistes de la communication et des techniques.

De manière un peu schématique, on peut identifier trois modalités, plus ou moins réflexives et contrôlées, de mobilisation des données du Web pour produire des énoncés sociologiques : la mesure des « effets du Web », l'usage d'Internet comme dispositif expérimental, et enfin une vision des espaces sociaux numériques comme homothétiques des espaces hors ligne.

2.1. Les effets du Web

Une première posture consiste à mobiliser les données du Web pour évaluer les « effets » du numérique sur un secteur d'activité. Prenant acte de l'intrication croissante entre nos activités en ligne et hors ligne, ces travaux rapportent l'activité en ligne autour des personnes et des objets à l'activité qu'ils suscitent hors ligne, dessinant des relations de dépendance, de cause et d'effet entre les scènes sociales. Ce type de démarche est particulièrement mobilisé en économie et en sciences du marketing.

Les recherches sur les systèmes de notes et avis en ligne en fournissent un bon exemple. Ces travaux s'attachent à comprendre le fonctionnement des évaluations laissées par les internautes sur les produits. Ils produisent ainsi, d'une part, une analyse du fonctionnement de cette évaluation profane, montrant par exemple qu'elle se concentre sur les produits stars et sur les produits de niche, et tend à négliger les produits de popularité intermédiaire ; et qu'elle est, dans l'ensemble, très clémente, les mauvaises notes étant minoritaires. D'autre part, ces travaux mettent en regard les évaluations en ligne avec des données d'évaluation hors ligne, et avec des indicateurs de succès et de vente, afin de mesurer les effets de la critique profane numérique sur les marchés concernés. Les chercheurs constatent ainsi un effet positif du volume d'avis en ligne sur les ventes de livres (Chevalier et Mayzlin, 2006), sur les entrées au cinéma (Liu, 2006 ; Larceneux, 2007), sur le chiffre d'affaires des restaurants (Luca, 2011) ; les constats sont plus nuancés quant aux effets de la valence des notes, certains chercheurs estimant que seul le nombre d'avis, par la visibilité qu'il procure, a un effet. Mobilisant des outils de *text mining*, Ghose et Impériotis (2011) estiment que les avis les plus explicitement subjectifs ont un effet fortement positif sur les ventes, tandis que les avis plus objectifs sont à la fois plus appréciés et plus défavorables aux ventes : les internautes voient leur intention d'achat confirmée par l'enthousiasme subjectif, mais elle est ralentie par les remarques plus objectives. En économie de la culture, où ces données ont été mobilisées de façon relativement précoce, il existe de nombreux débats économétriques pour affiner le sens des causalités entre le marketing hors ligne, le bouche à oreille en ligne et le succès commercial d'un produit[3]. Asur et Huberman (2010) utilisent par exemple avec succès le rythme des tweets concernant un film pour prédire sa réussite commerciale ; ils notent cependant qu'il s'agit moins de causalité que de covariation, les tweets comme les entrées en salle étant influencés par l'intensité du marketing. Dans une perspective différente, nos travaux sur les notes et avis dans le secteur de la restauration montrent comment ceux-ci prolongent un double mouvement de démocratisation du marché, en étendant le domaine de l'évaluation à des restaurants auparavant ignorés, et en redistribuant (de façon certes encore inégale) la participation à l'évaluation (Mellet *et al.*, 2014).

Dans un autre registre, les données du Web alimentent la question de l'évolution des sociabilités et de la circulation de l'information dans un contexte de diffusion des outils socionumériques. Un exercice récurrent des chercheurs du Web consiste ainsi à mesurer, sur les différents réseaux sociaux, l'évolution du nombre de contacts des individus, afin de

3. Pour une synthèse, voir Beuscart et Mellet (2012), p. 16-26.

(re)discuter la thèse de Robert Putman d'une diminution de l'intensité de la sociabilité ; et d'évaluer la distance moyenne entre deux personnes, dans la lignée des « six degrés de séparation » théorisés par Stanley Milgram (Ugander *et al.*, 2011). D'autres mesurent la circulation et la transformation des contenus, tels que les blagues, pour estimer l'impact d'Internet sur la mondialisation, au sens ici de circulation des idées ; ils observent ainsi une différence entre des blagues faisant l'objet d'une circulation globale, au prix d'une traduction et d'aménagements culturels relativement restreints, et des blagues locales résistantes à l'exportation. Les auteurs concluent néanmoins que « les blagues Internet fonctionnent comme un puissant (bien que souvent invisible) agent de mondialisation et d'américanisation » (Shifman, Levy et Thelwall, 2014). D'autres travaux s'efforcent de comprendre selon quelles modalités les espaces publics en ligne reconfigurent le débat et la conversation politique locale (Parasie et Cointet, 2012) ou globale (Wojcik, 2011).

2.2. Le Web comme dispositif expérimental

Une deuxième posture repose sur la mobilisation du Web comme dispositif expérimental. Les données, même massives, sont alors au moins partiellement construites par le protocole d'enquête, et complétées par des traces d'activité non suscitées par le sociologue.

Un dispositif pionnier et exemplaire de ce type a été construit par Salganik, Dodds et Watts (2006) pour étudier le rôle de l'information sociale dans la consommation culturelle. Les auteurs mettent en place un site *ad hoc* proposant 48 chansons gratuites, enregistrées par des groupes très peu connus. Ils sollicitent les internautes pour venir simplement écouter et télécharger des chansons ; pour être téléchargée, une chanson doit être préalablement écoutée en entier, de manière à s'assurer que le téléchargement témoigne d'une appétence pour la chanson. Cette expérience naturelle sur l'expression des goûts musicaux dans la consommation, qui neutralise autant que faire se peut les effets de notoriété antérieure des artistes et des œuvres, produits deux résultats. Dans un premier temps, les internautes (n = 14 341) participant à l'enquête sont répartis aléatoirement dans plusieurs échantillons, et les données montrent que le classement des chansons les plus appréciées est extrêmement varié d'un groupe à l'autre. Cela reflète certes l'hétérogénéité des goûts des différents groupes (même s'ils sont plutôt homogènes en termes de recrutement, des étudiants et jeunes internautes), mais aussi la forte dispersion de la consommation dans une situation – artificielle – d'absence d'information sur les produits. Ce résultat, produit de la situation expérimentale, est à mettre en regard de la distribution typique de la consommation des biens culturels, où l'essentiel de la demande se porte sur une petite partie des biens. Dans un second temps, les auteurs du dispositif ont commencé à afficher les compteurs de téléchargement des œuvres dans une moitié seulement des échantillons. Rapidement, les internautes bénéficiant de l'affichage ont commencé à avoir des comportements beaucoup plus moutonniers, et les échantillons avec information sociale à produire une courbe de distribution de la consommation bien plus classique des marchés culturels. Les auteurs se mettent ainsi en situation de mesurer l'effet pur de « l'information sociale », terme par lequel la théorie économique désigne l'observation des comportements des autres consom-

mateurs sur les marchés. Ils poussent l'expérimentation jusqu'à falsifier les compteurs pour quelques échantillons : le morceau le mieux classé devient dernier, le second avant-dernier, etc. Ils observent que, sauf pour le premier-devenu-dernier qui remonte la pente, les effets de l'inversion sont durables : les effets de l'information sociale l'emportent sur ceux de l'évaluation directe des biens par les consommateurs. Les données de cette expérience, partagées avec la communauté des chercheurs, ont été retravaillées par la suite pour affiner les interprétations (Krumme *et al.*, 2012).

Dans une autre recherche, Goel, Mason et Watts (2010) s'appuient sur les réseaux sociaux en ligne pour explorer les dimensions de la socialisation politique et de la construction de l'opinion publique. Les auteurs se situent par rapport au débat sur la polarisation croissante des opinions aux États-Unis, souvent affirmée mais difficile à vérifier empiriquement. Watts et son équipe s'appuient sur Facebook, conçu comme une explicitation du réseau social des individus, et recueillent auprès d'un large échantillon (n = 2500) des données comportant : les opinions de A sur le sujet X ; les opinions de B (ami de A) sur X ; les opinions de A sur l'opinion de B sur X. Pour ce faire, ils ont développé une application Facebook (FriendSense). Ils ont obtenu en fin de compte 1200 dyades complètes (opinions et perceptions). Sans surprise, ils vérifient que les opinions des amis sont plus proches que celle des étrangers : les individus ont tendance à se lier à des gens d'opinions similaires. Mais elles ne sont souvent pas si similaires qu'ils le croient : l'homophilie perçue est bien supérieure à l'homophilie réelle. En cas de désaccord, seuls 40 % des individus sont conscients de ce désaccord, tandis que les 60 % restants estiment à tort que leurs amis sont d'accord avec eux. Le questionnaire appuyé sur le réseau social en ligne permet la constitution de données originales à moindre coût.

Dans le même esprit, un nombre croissant de travaux s'appuient sur Facebook pour recueillir les traces d'activité d'individus, interrogés par ailleurs de façon classique sur leurs préférences. Castilho *et al.* (2014), s'intéressant à la dynamique de formation des groupes de travail, interrogent des étudiants sur ceux de leurs congénères avec lesquels ils aimeraient travailler pour les projets scolaires, et recherchent dans leurs interactions passées sur Facebook des prédicteurs des associations (sans surprise, les liens d'affinité sont de meilleures explications de la formation des groupes que les notes scolaires). Un autre exemple de la combinaison de données Web et de questionnaire est l'étude de Gomez Rodriguez *et al.* (2014) sur la surcharge informationnelle, dans laquelle ils combinent des données d'enquête quantitative auprès d'utilisateurs de Twitter avec des observations de leurs comportements passés sur le site, pour décrire la façon dont ils mettent en place des routines et des systèmes de traitement de l'information.

2.3. Le Web comme représentation homothétique de la société

Une dernière catégorie, plus hétérogène, regroupe les travaux construisant des énoncés généraux sur la société à partir des données du Web. La généralité du constat ou de la théorie est plus ou moins étoffée épistémologiquement selon les auteurs ; certains travaux peuvent être taxés de positivisme, ou faire l'objet des reproches de non-représentativité

formulés par Boyd et Crawford (2012) ou Tufekci (2014), les chercheurs semblant supposer que les comportements qu'ils observent en ligne seront bientôt diffusés à l'ensemble de la population. Cela n'empêche pas que les données en ligne fournissent des éclairages inédits sur des questions auparavant difficiles à investiguer empiriquement, éclairages dont il reste à trouver les conditions de généralisation.

Les travaux conduits par l'équipe de recherche de Facebook, Facebook Data Science, sont assez représentatifs de cette hésitation sur la portée à donner aux résultats – empiriquement inédits – construits à partir des données en ligne, en l'occurrence les données massives du premier site de réseau social. Dans un article de 2012, Bashky *et al.* mettent ainsi en place un protocole visant à mesurer l'influence des liens forts et des liens faibles dans la diffusion de l'information. À partir d'un « échantillon » très conséquent (253 millions d'utilisateurs de Facebook), ils vérifient tout d'abord qu'il y a une probabilité plus grande pour des individus de partager un contenu si celui-ci a été partagé par leurs amis (même s'ils ont pu y être exposés par ailleurs) ; ensuite, que la probabilité de diffuser un contenu qui a été partagé par un lien fort est plus forte, toute choses égales par ailleurs, que si celui-ci a été partagé par un lien faible – la force du lien étant mesurée par le nombre d'échanges sur le site. C'est néanmoins un troisième constat qui est placé au cœur de l'article : les liens faibles rendent visibles des contenus et des informations qui seraient sinon restés invisibles à l'internaute. La probabilité pour un contenu d'être remarqué augmente d'un facteur de 23 s'il est posté par un lien faible, car il s'agit d'une information qui n'entre pas dans les sources habituelles de l'individu. Cet accroissement n'est que d'un facteur 7 dans le cas des informations postées par les liens forts, du fait de l'homophilie (les liens forts ayant des univers informationnels plus proches). Il est intéressant d'observer la façon dont les auteurs concluent leur démonstration. Dans un premier temps, ils présentent leurs résultats comme une généralisation des travaux classiques de Granovetter sur la compréhension de la diffusion ordinaire de l'information, confirmant à une grande échelle et pour une large variété de contenus informationnels la capacité supérieure des liens faibles à apporter de l'information inédite. Ils restreignent ensuite le champ des conclusions à l'analyse des grands réseaux sociaux en ligne, en estimant que la diffusion de l'information y est plus fluide, moins enclavée au sein de cliques, du fait du plus grand nombre de liens faibles et de leur meilleure distribution (notons que ce résultat est convergent avec les objectifs de communication de l'entreprise Facebook, qui se présente comme un facteur de fluidité sociale). Les auteurs émettent pour terminer l'hypothèse que l'adoption en masse des réseaux socio-numériques transformera de façon plus générale les formes d'exposition des individus à l'information, et que les constats – optimistes – faits sur « l'échantillon » des utilisateurs de Facebook auront bientôt une valeur plus générale. C'est donc par la supposition d'une diffusion des usages numériques à l'ensemble de la population, diffusion supposée uniforme et homogène, que les auteurs produisent un constat social général.

La même oscillation peut s'observer dans les travaux sur les goûts construits à partir des données du Web. McAuley et Leskovec (2013) s'appuient ainsi sur les données exhaustives de plusieurs sites de notation de biens culturels (au sens large : films, bières, vins, restaurants), et analysent les trajectoires de notation – donc de goût – des

participants à ces sites. Ils observent que, dans l'ensemble, ces trajectoires sont très similaires, et conduisent d'un statut d'« amateur », qui valorise des biens d'accès facile (par exemple les bières blondes de marque commerciale), à celui de « connaisseur », qui apprécie des biens d'accès plus difficile (par exemple les bières brunes artisanales). Symétriquement, se dessine un espace des biens très nettement hiérarchisé. Autrement dit, au prix d'un travail minutieux de reconstitution et de comparaison des trajectoires individuelles très hétérogènes, les auteurs soutiennent non seulement qu'il existe sur ces sites une éducation au goût, à laquelle personne n'échappe sinon par l'abandon du site, mais que le bon goût vers lequel tend cette éducation est unique et homogène (et non pas éclaté entre des goûts individuels irréductibles). Ici, les auteurs, ancrés dans la discipline informatique, concluent en imaginant des améliorations des systèmes de recommandation, laissant aux sciences humaines le soin de mesurer les possibilités de généraliser ce constat au-delà de l'univers de l'évaluation en ligne, et de le confronter aux théories de la distinction et de l'omnivorisme.

Cette extrapolation vers le hors-ligne, qui omet à la fois de discuter du biais d'échantillonnage et de l'équivalence entre comportement en ligne et hors ligne, est relativement fréquente dans les travaux des sciences informatiques ; ainsi le travail de Silva *et al.* (2014) analyse les distances culturelles à partir des habitudes culinaires des individus, elles-mêmes mesurées à partir des *check in* dans les restaurants sur Foursquare. Inversement, certains travaux peuvent présenter un intérêt sociologique évident sans avoir pourtant cette prétention. San Pedro *et al.* (2012), dans un travail de recherche informatique visant à améliorer les systèmes de recherche d'image, analysent les données d'un très vaste concours de photographie, où les images ont été évaluées et commentées par des photographes amateurs. Leur analyse textuelle ressort un ensemble d'une trentaine de mots les plus souvent mobilisés pour évaluer les œuvres, et qui pourraient parfaitement décrire les conventions d'évaluation du monde de l'art de la photographie amateur.

Il serait malvenu, au regard de leur virtuosité de leur inventivité méthodologique, de disqualifier ces travaux de *webscience* au motif d'un manque de rigueur dans l'écriture de leurs conclusions et dans la définition de la portée de leurs énoncés sur la société. Ils sont souvent inscrits dans le champ disciplinaire des sciences informatiques, et n'ont pas nécessairement pour objet premier la contribution à la théorie sociologique. Il convient d'éviter, au motif de critiques méthodologiques légitimes, de jeter le bébé avec l'eau du bain, en taxant les travaux issus des données du Web de « réductionnisme tautologique » (Denouël et Granjon, 2011), et plutôt de souligner les potentialités d'un dialogue interdisciplinaire aujourd'hui encore émergent. On observe ce dialogue dans le développement des conférences comme ICWSM, dans la constitution d'unités de recherches (comme l'Institut des systèmes complexes), dans la circulation des chercheurs (Duncan Watts, physicien de formation, occupe une chaire de sociologie à Columbia) et des méthodes. Espérons que cette circulation permettra, en même temps que la diffusion des innovations méthodologiques, leur réglage épistémologique.

Conclusion

Du fait de leur mode de construction, les données du Web sont affectées de plusieurs faiblesses pour l'analyse sociologique, notamment une représentativité difficile à établir et un manque d'épaisseur sociologique des acteurs et des actions. Pour autant, comme nous espérons l'avoir montré à travers la multiplication des exemples, elles constituent un matériau précieux pour l'enquête. Par leur volume, elles permettent d'appréhender de larges espaces sociaux, tout en restituant la diversité ; produites en situation, elles sont exonérées de certains biais déclaratifs des méthodes classiques de la sociologie, questionnaires et entretiens, avec lesquelles elles peuvent être combinées de façon très profitable. En outre, le traitement de ces données est l'occasion d'un dialogue avec les sciences informatiques, susceptible de renouveler et d'enrichir les formes de manipulation, de représentation et d'interprétation des données, et *in fine* l'analyse sociologique.

Les données alimentent ainsi la compréhension des pratiques sociales en ligne, enrichissant la sociologie des nouveaux média, la compréhension des engagements en ligne, des sociabilités, des espaces publics sur Internet, etc. Mais elles fournissent aussi, plus généralement, des visions renouvelées sur les pratiques sociales hors ligne, qui sont de plus en plus intriquées avec les pratiques connectées dont les données fournissent une vue. Il reste à mener de façon plus systématique une réflexion sur les conditions d'extrapolation des constats, qui reste encore embryonnaire et située.

Références

Adamic L. et Huberman B.A. (2000), « Power-law distribution of the World Wide Web », *Science*, vol. 287, p. 2115a.

Adamic L. et Glance N. (2005) « The political blogosphere and the 2004 U.S. election: divided they blog », *Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD2005*, ACM, p. 36-45.

Asur S. et Huberman B. (2010), « Predicting the future with social media », *International Conference on Web Intelligence and Intelligent Agent Technology* (Toronto), ACM.

Bakshy E., Rosenn I., Marlow C., Adamic L., Park M. et Arbor A. (2012), « The role of social networks in information diffusion », *Proceedings of the 21st international conference on World Wide Web* (Lyon), ACM.

Barabasi A.L., Ravasz E. et Vicsek T. (2001), « Deterministic scale-free networks », *Physica*, vol. 299, nº 3, p. 559-564.

Bastard I., Bourreau M., Maillard S., Moreau F. (2012), « De la visibilité à l'attention : les musiciens sur Internet », *Réseaux*, nº 175, p. 19-42 (en ligne : https://www.cairn.info/revue-reseaux-2012-5-page-19.htm).

Beauvisage T., Beuscart J.-S., Couronné T. et Mellet K. (2011), « Le succès sur Internet repose-t-il sur la contagion ? Une analyse des recherches sur la viralité », *Tracés*, nº 21, p. 151-166 (en ligne : https://traces.revues.org/5194).

Beauvisage T. (2013), « Compter, mesurer et observer les usages du web : outils et méthodes », *in* Barats C., *Manuel d'analyse du web en sciences sociales*, Paris, Armand Colin.

Beuscart J.-S. (2008), « Sociabilité, notoriété virtuelle et carrière artistique. Les musiciens autoproduits sur MySpace », *Réseaux*, nº 152, p. 139-168 (en ligne : https://www.cairn.info/revue-reseaux1-2008-6-page-139.htm).

Beuscart J.-S. et Couronné T. (2009), « La distribution de la notoriété en ligne. Une étude quantitative de MySpace », *Terrains et travaux*, nº 15, p. 147-170 (en ligne : https://www.cairn.info/revue-terrains-et-travaux-2009-1-page-147.htm).

Beuscart J.-S., Cardon D., Pissard N., Prieur C. et Pons P. (2009), « Pourquoi partager mes photos de vacances avec des inconnus ? Une étude de Flickr », *Réseaux*, nº 154, p. 91-129 (en ligne : https://www.cairn.info/revue-reseaux-2009-2-page-91.htm).

Beuscart J.-S. et Beauvisage T. (2012), « Audience dynamics of online catch up TV », *Proceeding of the 21st international conference on World Wide Web* (Lyon), ACM.

Beuscart J.-S. et Mellet K. (2012), *Promouvoir les œuvres culturelles. Usages et efficacité de la publicité dans les filières culturelles*, Paris, La Documentation française.

Beuscart J.-S. et Crepel M. (2014), « Les plateformes d'autopublication artistique en ligne. 4 figures de l'engagement des amateurs dans le Web 2.0. », *in* Lizé W., Naudier D. et Sofio S., *Les Stratèges de la célébrité. Intermédiation et consécration dans les univers artistiques*, Paris, La Documentation française.

Blank G. et Reisdorf B. (2012), « The participatory web: A user perspective on Web 2.0 », *Information, Communication and Society*, vol. 15, nº 4, p. 537-554.

Blank G. (2013), « Who creates content? Stratification and content creation on the Internet », *Information, Communication and Society*, vol. 16, nº 4, p. 590-612.

Boyd D. et Crawford K. (2012), « Critical questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon », *Information, Communication and Society*, vol. 15, n° 5, p. 662-679.

Cardon D., Roth C. et Fouetillou G. (2011), « Two paths of glory-structural positions and trajectories of websites within their topical territory », *International Conference on Weblogs and Social Media* (Barcelone), AAAI, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2765 (dernière consultation le 8 mars 2017).

Castilho D., Vaz de Melo P., Querciay D. et Benevenuto F. (2014), « Working with friends: Unveiling working affinity features from Facebook data », *International Conference on Weblogs and Social Media* (Ann Arbor), AAAI, https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8084.

Caverlee J. et Webb S. (2008), « A large-scale study of MySpace: Observations and implications for online social networks », *International Conference on Weblogs and Social Media* (Seattle), AAAI.

Cha M., Kwak H., Rodriguez P., Ahn Y. et Moon S. (2007), « I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system », *IMC'07* (San Diego).

Cha M., Mislove A. et Gummadi K. (2009), « A measurement-driven analysis of information propagation in the Flickr social network », *Proceedings of the 18th International Conference on World Wide Web WWW'09* (Madrid), ACM.

Cha M., Mislove A. et Gummadi K. (2008), « Characterizing social cascades in Flickr », *WOSN'08* (Seattle), ACM.

Cha M., Haddadi H., Benevenuto F. et Gummadi K. (2010), « Measuring user influence in Twitter: The million follower fallacy », *International Conference on Weblogs and Social Media* (Washington), AAAI, https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8084.

Chevalier J. et Mayzlin D. (2006), « The effect of word of mouth on sales: Online book reviews », *Journal of Marketing Research*, vol. 43, n° 3, p. 345-354.

Crépel M. (2011), *Tagging et folksonomies : pragmatique de l'orientation sur le Web*, thèse de doctorat, Université de Rennes 2 (http://tel.archives-ouvertes.fr/tel-00650319).

Denouël J. et Granjon F. (2011), « Penser les usages sociaux des technologies numériques d'information et de communication », *in* Denouël J. et Granjon F. (dir.), *Communiquer à l'ère numérique*, Paris, Presses des Mines.

Friggeri A., Adamic L., Eckles D. et Cheng J. (2014), « Rumor Cascades », International Conference on Weblogs and Social Media (Ann Arbor), AAAI, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8122 (dernière consultation le 8 mars 2017).

Ghose A. et Impeirotis P. (2011), « Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer charactéristics », *IEEE Transactions on Knowledge and data engineering*, vol. 23, n° 10.

Godes D. et Mayzlin D. (2009), « Frim-created word-of-mouth communication: Evidence from a field test », *Marketing Science,* vol. 28, n° 4, p. 721-739.

Goel S., Mason W. et Watts D.J. (2010), « Real and perceived attitude agreement in social networks », *Journal of Personality and Social Psychology*, vol. 99, n° 4, p. 611.

Gomez Rodriguez M., Gummadi K. et Schoelkopf B. (2014), « Quantifying information overload in social media and its impact on social contagions », *International Conference on Weblogs and Social Media* (Ann Arbor), AAAI, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8108 (dernière consultation le 30 mars 2017).

Herring S.C., Kouper I., Paolillo J.C., Scheidt L.A., Tyworth M., Welsch P., Wright E. et Yu N. (2005), « Conversations in the blogosphere: An analysis "from the bottom up" », *Proceedings of the Thirty-Eighth Hawai'i International Conference on System Sciences HICSS-38* (Los Alamitos), IEEE Press (en ligne: http://ella.slis.indiana.edu/~herring/blogconv.pdf, dernière consultation le 30 mars 2017).

Huberman B.A., Romero D.M. et Wu F. (2009), « Crowdsourcing, attention and productivity », *Journal of Information Science*, vol. 35, nº 6, p. 758-765.

Jouët J. (2011), « Des usages de la télématique aux *Internet Studies* », *in* Denouël J. et Granjon F. (dir.), *Communiquer à l'ère numérique*, Paris, Presses des Mines.

Katz E. et Lazarsfeld P. (1955), *Personal Influence: The Part Played by People in the Flow of Mass Communications*, New York, The Free Press ; traduction française : *Influence personnelle. Ce que les gens font des médias*, Paris, Armand Colin, 2008.

Krumme K., Cebrian M., Pickard G. et Pentland A. (2012), « Quantifying social influence in an online cultural market », *PLOS One*, vol. 7, nº 5, e33785.

Kumar R., Novak J. et Tomkins A. (2006), « Structure and evolution of online social networks », *KDD'06 Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (Philadelphie), ACM.

Larceneux F. (2007), « Buzz et recommandations sur internet : quels effets sur le box-office ? », *Recherche et applications en marketing*, vol. 22, nº 3, p. 45-64.

Lehmann J., Goncalves B., Ramasco J. et Cattuto C. (2012), « Dynamical classes of collective attention on Twitter », *Proceeding of the 21st International Conference on World Wide Web* (Lyon), ACM.

Leskovec J., Adamic L. et Huberman B. (2006), « The dynamics of viral marketing », *Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06)*, ACM.

Leskovec J., Backstrom L. et Kleinberg J. (2009), « Meme-tracking and the dynamics of the news cycle », *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, p. 497-506.

Levrel J. (2006), « Wikipedia, un dispositif médiatique de publics participants », *Réseaux*, nº 136, p. 185-218 (en ligne : https://www.cairn.info/revue-reseaux1-2006-4-page-185.htm).

Liu Y. (2006), « Word of mouth for movies: Its dynamics and impact on box office revenue », *Journal of Marketing*, vol. 70, nº 3, p. 74-89.

Liu Y., Kliman-Silver C. et Mislove A. (2014), « The tweets they are a-changin': Evolution of Twitter users and behavior », *International Conference on Weblogs and Social Media* (Ann Arbor), AAAI, www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8043 (dernière consultation le 30 mars 2017).

Luca M. (2011), *Reviews, reputation, and revenue: The case of Yelp.com*, Harvard Business School Working Paper, nº 12-016.

McAuley J.J. et Leskovec J. (2013), « From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews », *Proceedings of the 22nd International Conference on World Wide* Web (Rio de Janeiro), New York, ACM.

Michael L. et Otterbacher J. (2014), « Write like I write: Herding in the language of online reviews », *International Conference on Weblogs and Social Media* (Ann Arbor), AAAI, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8046.

Mellet K, Beauvisage T., Beuscart J.-S. et Trespeuch M. (2014), « A democratization of markets? Online consumer reviews in the restaurant industry », *Valuation Studies*, nº 2.

Mislove A., Gummadi K., Druschel P. et Bhattacharjee B. (2007), « Measurement and analysis of online social networks », *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement IMC'07* (San Diego), ACM.

Parasie S. et Cointet J.-P. (2012), « La presse en ligne au service de la démocratie locale », *Revue française de science politique*, vol. 62, p. 45-70.

Prieur C., Cardon D., Beuscart J.-S., Pissard N. et Pons P. (2008), « The strenght of weak cooperation: A case study on Flickr », arXiv.org, arXiv:0802.2317.

Salganik M.J., Dodds P.S. et Watts D.J. (2006), « Experimental study of inequality and unpredictability in an artificial cultural market », *Science*, vol. 311, p. 854-856.

San Pedro J., Yeh T. et Oliver N. (2012), « Leveraging user comments for aesthetic aware image search reranking », *Proceeding of the 21st international conference on World Wide Web* (Lyon), ACM.

Shifman L., Levy H., Thelwall M. (2014), « Internet jokes: The secret agents of globalization? », *Journal of Computer Mediated Communication*, vol. 19, nº 4, p. 727-743.

Silva T.H., Vaz de Melo P., Almeida J., Musolesi M. et Loureiro A. (2014), « You are what you eat (and drink): Identifying cultural boundaries by analyzing food and drink habits in foursquare », *Proceedings of the Eighth International Conference on Weblogs and Social Media* (Ann Arbor), AAAI, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8113.

Stoica A., Couronné T. et Beuscart J.-S. (2010), « To be a star is not only metaphoric: From popularity to social linkage », *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (Washington), AAAI, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1480.

Tufekci A. (2014), « Big questions for social media Big Data: Representativeness, validity and other methodological pitfalls », *Eighth International Conference on Weblogs and Social Media* (Ann Arbor), 2014, AAAI, www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062.

Ugander J., Karrer B., Backstrom L. et Marlow C. (2011), « The Anatomy of the Facebook Social Graph », arXiv:1111.4503.

Watts D.J. et Dodds P.S. (2007), « Influentials, networks, and public opinion formation », *Journal of Consumer Research*, vol. 34, nº 4, p. 441-458.

Wojcik S. (2011), « Prendre au sérieux la démocratie électronique. De quelques enjeux et controverses sur la participation politique en ligne », *in* Forey E. et Geslot C. (dir.), *Internet, machines à voter, démocratie*, Paris, L'Harmattan.

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 circulaires (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.

Postface

« Un travail de fourmi »

Simon Paye
Sociologue, maître de conférences à l'université de Lorraine

L'utilisation intensive des algorithmes dans l'industrie du numérique semble dessiner les contours d'un modèle de création de valeur sans intervention humaine. L'automatisme serait même poussé à son paroxysme puisque, contrairement à des robots sur une ligne de fabrication industrielle, les algorithmes qui travaillent la matière numérique sont adaptatifs et peuvent redéfinir leurs consignes initiales. Mais c'est précisément lorsque le mythe du « tout automatisé » semble se réaliser qu'il se montre irréaliste. Car les machines, même apprenantes, laissent toujours vierge un territoire de tâches et d'activités qu'elles ne sauraient faire. Cette irréductibilité du travail humain est patente dans le modèle économique des entreprises du numérique. Leur chaîne de création de valeur fait intervenir, en amont, un travail de conception et de reparamétrage des algorithmes (Cardon, 2013). Ensuite, le traitement algorithmique de gigantesques flux de données a pour carburant le « travail gratuit » des utilisateurs d'objets connectés qui cèdent leurs traces numériques[1]. Enfin, en aval, des milliers de travailleurs sont mobilisés, soit pour combler les territoires de tâches et d'activités non automatisées, soit pour évaluer l'efficacité de nouvelles versions d'algorithmes à tester. Salariés ou indépendants, ils œuvrent dans les coulisses de l'industrie de la donnée et contribuent, autant que les algorithmes, à la prospérité des entreprises du numérique.

L'une des priorités pour les sciences sociales du travail est sans conteste de donner la parole à tous les types de « travailleurs de la donnée » (Bastard *et al.*, 2013), tant pour saisir les innombrables actes de travail qu'exigent au quotidien la production et l'exploitation des données numériques, que pour comprendre le sens donné à ce travail par ceux et celles qui le font.

L'entretien qui suit, mené en avril 2015 avec Émilien[2], un *rater* (évaluateur) de la région parisienne, permet d'aborder ces questions[3]. Les *raters* constituent une figure à la fois emblématique et peu connue du travail dans l'industrie de la donnée. On ne sait pas combien ils sont, encore moins qui ils sont. Étudiants, *homeworkers*, femmes au foyer, pré-

1. Voir le chapitre 3 de ce volume pour une discussion du qualificatif de « travail » concernant la production de traces numériques exploitables par les entreprises du numérique.
2. Le prénom de l'enquêté a été modifié.
3. Ces questions sont également traitées dans le chapitre 6 de cet ouvrage, qui donne la parole quant à lui aux travailleurs intervenant dans les coulisses de l'Open Data.

caires des cinq continents, ils travaillent indirectement pour Google, Amazon ou Microsoft. Ils s'appellent eux-mêmes *human raters* quand le langage plus officiel des intitulés les baptise *internet evaluators* ou *internet assessors*. On connaît peu les entreprises de sous-traitance qui se chargent de rémunérer leurs prestations et d'organiser une vente de travail de masse, comptabilisé à la minute. Enfin, on ne sait pas toujours ce que font précisément ces *raters*, si ce n'est qu'ils travaillent en ligne sur leur propre outil de travail (un ordinateur ou un *smartphone*) et qu'ils sont payés pour compléter, faciliter ou évaluer le « travail » des algorithmes[4].

Notre entretien, dont le style informel a été conservé, a permis de saisir ce travail humain sous un angle moins promotionnel que ne le suggère l'expérience d'« analyste de performances » consignée dans le CV d'Émilien. Notre conversation d'environ deux heures[5] est revenue dans un premier temps sur une expérience de travail antérieure, dans un centre d'appel d'Apple en Irlande, puis a abordé plus longuement le quotidien du travail en tant que *rater* en sous-traitance pour Google.

Le parcours social d'Émilien

Émilien a 25 ans au moment de l'entretien. Son père, entrepreneur dans l'informatique, l'initie précocement à l'usage des ordinateurs. Sa mère travaille dans l'immobilier et touche un salaire modeste. Il grandit avec ses trois sœurs dans l'ouest de la région parisienne. Après un bac technologique, il obtient un DUT en réseaux et télécommunications à quelques kilomètres du foyer familial. Après cela, il s'inscrit en licence de langues orientales pour apprendre le chinois en attendant d'avoir les idées plus claires sur son orientation professionnelle. Il dépose quand même son CV sur le site de l'APEC[6]. Quelques mois plus tard, il est sollicité par Apple pour devenir téléconseiller en Irlande. Il y reste un an et demi et participe à un projet pilote de télétravail. Il travaille ainsi trois mois à domicile, puis quitte Apple pour revenir en France, où il redevient étudiant, cette fois en licence de sciences humaines et sociales. Il vit pendant environ deux ans sur ses économies, en logeant à peu de frais chez sa mère, alors divorcée. Pour sa troisième année de licence, les économies faisant défaut, il cherche un emploi à temps partiel. À la rentrée 2014, il découvre au gré de recherches sur Internet l'entreprise SuccessTech[7]. Celle-ci l'« embauche » sur un statut d'auto-entrepreneur pour travailler pour Google comme *internet assessor.*

4. Quelques indications d'ordre technique sont données par l'entreprise Google dans une vidéo intitulée « How does Google use human raters in web search? », https://www.youtube.com/watch ?v =nmo3z8pHX1E.
5. Quelques rares passages de l'entretien ont été retirés, soit pour préserver l'anonymat, soit pour éviter les redondances et les développements trop éloignés du sujet traité.
6. Association pour l'emploi des cadres.
7. Le nom de l'entreprise a été modifié.

1. L'entretien

1.1. Le point zéro du travail : « après que les algorithmes soient passés »

Je fais une recherche sur l'industrie des données, et une des questions qui m'interpellent, c'est la place du travail humain dans cette économie des big data*. Je me suis dit qu'une des utilisations possibles de ton témoignage de* Google rater *serait d'alimenter cette recherche sous la forme d'un entretien vivant. Et je pense qu'il faut respecter la règle de l'anonymat, c'est important.*

Moi, c'est pas que ça me dérangerait de parler en mon nom. Après, je ne sais pas si tu veux qu'on parle du boulot ou plus largement du sujet. Parce que le boulot, en principe, j'ai pas le droit. Donc du coup, oui, je préfère être anonymisé.

Moi ce qui m'intéresse... J'ai compris que tu bossais comme rater...

C'est ça, exactement.

Un des sujets que je cherche à approfondir, c'est : où se loge le travail humain dans l'économie des big data *? Quelles sont les formes du travail dans cette économie ?*

Ils ont un terme chez Google, je sais qu'ils emploient le terme *Hit*, *human intelligence tasks*, pour désigner le travail de complément qu'on fait après que les algorithmes soient passés.

1.2. L'expérience irlandaise : une socialisation progressive au travail en ligne

J'ai vu dans ton CV que tu avais travaillé comme téléopérateur en Irlande.

À Cork, si tu veux travailler chez Amazon ou Apple, tu peux y aller sans avoir de taf qui t'attend. Si tu connais des gens qui travaillent à l'intérieur, tu peux être pris très rapidement. Je le sais d'expérience parce que j'ai fait rentrer plusieurs personnes chez Amazon. Tu connais quelqu'un à l'intérieur, c'est des boîtes, ça respire. Leur DRH, ils vont virer 100 personnes dans la semaine, et ils vont en réengager 400 la semaine suivante, donc tu peux facilement faire entrer des gens.

Concrètement toi, comment ça s'est passé ? Comment tu en es arrivé à travailler là-bas en Irlande ?

C'est pas de mon fait. J'ai fait un DUT dans la banlieue ouest de Paris, je l'ai eu. Après j'ai fait une licence de chinois, mais je savais pas quoi faire, je savais pas où j'allais. J'ai fait ça en attendant d'avoir une révélation sur ce que je veux faire dans la vie. Et j'ai fait ça 6 mois ou un peu moins, parce qu'en janvier, j'ai reçu un coup de fil un soir à 18 h, alors que moi je cherchais pas de boulot. Il me semble pas que j'avais mis mon CV en ligne. Je ne vois pas comment ils ont pu me trouver, mais bref, je reçois un coup de fil, une fille qui me parle en anglais, qui me dit « Bonjour, qui est le plus grand constructeur d'électronique au monde ? » Je dis « J'en sais rien, Samsung ? », et elle me

dit « Ah non, non, c'était pas ça la réponse, mais c'est pas grave, c'était juste histoire de casser la glace. Est-ce que vous voudriez travailler en Irlande pour un gros constructeur d'électronique ? » Je t'assure que ça m'est tombé dessus, vraiment *out of the blue* quoi. Bref, avec la fille on parle une heure, elle m'explique qu'Apple recherche du monde pour ce truc-là. Et ce que j'ai vécu chez Apple, le travail en lui-même, c'était du travail de téléopérateur, mais les conditions de travail n'avaient absolument rien à voir avec un *call center* normal. On était bien payés, hyper bien traités, avec des conditions matérielles de travail absolument exceptionnelles. T'es chez Apple, quoi, c'est le QG européen, les bâtiments sont énormes, t'as des salles de sport à disposition... Des trucs cons aussi tu vois, mais ta machine de travail, t'es administrateur de tout, tu fais ce que tu veux. Y'a aucun contrôle, tu fais ce que tu veux pendant la journée, du moment que tu fais le boulot. Tu vois, c'est le management à la californienne quoi. Du moment que tu fais le taf, tu fais ce que tu veux, tu joues à la balle dans les bureaux si tu veux. Et le premier truc qui m'a fait découvrir ça, c'est que la fille qui m'a appelé, elle m'a redonné un rendez-vous téléphonique pour aller plus loin, et quand elle m'a présenté le boulot et les conditions, elle m'a dit que si je disais oui, je commençais la semaine suivante, on me payait l'avion, on me filait 1 500 euros pour m'installer à Cork. Moi ça me paraissait trop beau, tu vois.

Tu en as parlé à des gens ?

Non, parce que j'aurais pu demander autour de moi, mais j'ai pas pris le temps de faire des recherches parce que moi je savais pas où j'allais dans la vie. J'ai ce truc-là qui tombe totalement du ciel, l'appel, l'Irlande, j'ai dit oui, une semaine après ils m'envoyaient les billets d'avion et j'y allais.

C'était un contrat à durée déterminée ?

Non, c'était un CDI dès le début. Donc t'arrives, t'as 4 semaines de formation. Donc pendant la formation t'es payé à la semaine comme ça t'as du *cash* qui tombe tout de suite. Bien que les conditions de travail soient très bonnes, il y avait quand même une grosse rotation. Et malgré tout, eux, ben ils investissement quand même assez lourdement dans les gens qu'ils font venir. Et ils ont pas peur de mettre des sous dans les employés pour les garder. Mais ça colle un peu avec la manière dont on nous demandait de tafer cela dit parce que c'était vraiment l'excellence ou rien. C'était abusé les objectifs qu'on avait. Mais en revanche, à côté de ça, on nous chouchoutait vraiment à mort. Et je pense que dans des *call centers* où t'es plus payé au lance-pierres, le *turn-over* doit être bien plus élevé encore.

C'est quoi une semaine de travail par exemple ?

Ben par exemple, t'avais trois *shifts* différents : *early*, *medium* et *late*. *Early*, tu commences à 6 h du mat, tu finis à h. *Medium*, c'est des horaires normaux, 9 h-17 h. Et puis *late*, tu commences à midi, tu finis à 2 h je crois. La journée de taf habituelle, c'est tu t'assois, tu mets ton casque sur les oreilles, tu te mets en ligne et ça commence. Et ils ont une assez bonne gestion de la file d'attente. Par là je veux dire que quand tu finis un appel dans les 10 secondes t'en as un autre. C'est hypertaxant nerveusement comme boulot.

« Taxant » ?

Taxing, ouais... C'est fatigant. Je confonds peut-être les mots, je sais pas ! Ce que je veux dire c'est que, nerveusement, c'est pas évident. Parce que les gens qui t'appellent, ils sont toujours forcément insatisfaits. Mais après, une fois que t'apprends les techniques pour *diffuse anger* [dissiper la colère] comme ils disent, calmer le client, c'est assez... Et puis tu vois, comme on avait une marge de manœuvre vachement élevée par rapport à ce qu'on pouvait dire aux clients. Alors que dans les autres *call centers* en général, t'es hyper calé sur ce que tu dois dire, ce que tu dois faire, eux ils ont vraiment une manière de faire qu'est différente. Du moment que t'atteins l'objectif et que le client est satisfait et qu'il va revenir dans un Apple Store, tous les moyens sont bons. Donc la journée de taf habituelle, c'était ça quoi, c'est... appel-appel-appel-appel.

Et le temps de travail était régulé ? Tu comptais tes heures ?

Non, non, le temps de travail c'était... je sais plus, 35 h, je sais plus exactement... Le truc, ce qui arrivait souvent, c'est que la fin de ton *shift* arrive, t'as un appel compliqué qui va durer 1 h, 1 h 30 (parce que les appels, c'est pas 10 secondes, c'est vraiment pas de la hotline habituelle quoi, c'est 10 min minimum et ça peut prendre 2h s'il faut), tu peux pas dire au client « je vais raccrocher, j'ai fini mon *shift* ». Mais pas de souci, soit ils te paient en heures sup', soit tu reviens deux heures plus tard le lendemain, y'a pas de problème. C'est un boulot qu'est pas agréable, mais l'employeur est super.

Tu avais des clients francophones ?

C'était en anglais et en français. En fait, au début, tu fais que les lignes françaises, et après les [lignes] anglaises ; ils ont un gros volume d'appels et quand ils ont de l'*overflow*, du surplus, quand il y a trop d'attente sur les piles, certains gars peuvent être mis sur les lignes anglaises, ceux qui se sentent à l'aise, hein. Parce que tu vois chez Apple, il est hors de question qu'un client appelle et tombe sur un mec qui a un très fort accent français et qui a du mal à comprendre.

Tu fais le reporting*, tu notes tout en même temps ?*

Voilà. *Real-time reporting*.

T'es resté combien de temps sur ce poste ?

En fait, au bout d'un peu plus d'un an, ils ont commencé à vouloir former des gens *at home* – des gens qui puissent travailler depuis chez eux – pour rendre leur force de travail plus flexible. Ça a été mon premier contact avec le monde du *crowdsourcing*, du travail à la maison. Ce qui s'est passé, c'est qu'il y a une cinquantaine de personnes qui ont été renvoyées chez eux. Si tu veux, on connaissait déjà le boulot, on nous faisait bosser depuis chez nous,

d'abord pour mettre en place certains aspects techniques et informatiques, et aussi pour faire un projet pilote et voir comment ça se passe. Et les trois derniers mois en Irlande, j'ai bossé depuis chez moi.

Comment ça s'appelait ce job ? Ils appelaient ça comment ?

Customer representative, un truc comme ça.

On te l'a proposé ? Comment ça a marché ?

En fait, le projet a été annoncé. Ils ont demandé qui serait intéressé. Moi j'ai proposé, ils ont retenu 50 personnes, et donc voilà, tout simplement quoi.

Et tu vivais dans quelles conditions ? En colocation ?

J'étais en colocation avec un pote que j'ai fait venir en Irlande pour lui trouver du boulot. Un pote qui galérait, je lui ai dit « Écoute, viens à Cork, je peux te faire manger, te loger, te trouver du boulot vite fait ». Il est venu et, en une semaine, il était pris chez Amazon.

Et t'arrivais à bosser en tant que representative *chez toi ?*

Ouais, c'est pas évident parce que ça a des bons côtés mais des mauvais aussi. Quand t'as du monde avec toi, évidemment, t'es beaucoup moins concentré sur le taf. Tu vas dire « OK, bon, je peux me réveiller plus tard ». Sauf que quand tu as le malheur de te rendormir et de te réveiller 5 secondes avant ton heure, tu sors de ton lit, tu te mets sur ta chaise, tu sors ton casque, et t'es à peine réveillé, y'a le client qui te dit bonjour et c'est horrible.

Donc t'as fait trois mois comme ça ?

À peu près oui. Après, j'ai quitté la boîte parce que j'ai décidé de reprendre mes études.

Tu parles de crowdsourcing *mais, en même temps, c'est du télétravail ? Quelle est la différence ? Comment ça marchait ce truc-là ? T'étais toujours employé ou est-ce que... ?*

Oui, quand j'étais chez moi j'étais toujours employé, parce que c'était juste un projet pilote pour voir ce que ça donnait de faire travailler les gens depuis chez eux. J'ai même pas changé de contrat. Je parle de *crowdsourcing*, mais c'est pas du *crowdsourcing;* le but de ce projet c'était, si ça marchait bien, de pouvoir lever des armées assez rapidement, et de simplement fournir un iMac à la personne chez lui, et c'est beaucoup plus rapide et beaucoup plus flexible pour eux d'engager des gens comme ça. Mais c'est pas vraiment du *crowdsourcing* au sens où ce que je fais pour Google en ce moment par exemple, là c'est vraiment du *crowdsourcing* pur et dur. C'est-à-dire que c'est pas vraiment Google qui m'emploie, c'est une boîte qui s'appelle SuccessTech, qui est spécialisée dans le *crowdsourcing*. Eux, ils lèvent des armées de gens dont ils vendent les minutes de travail à Google.

1.3. Devenir *Google rater*

Et à quel moment tu t'es mis à bosser pour Google ?

En fait, j'avais ramené pas mal de sous d'Irlande, et je m'étais dit que j'allais essayer de tenir le plus longtemps possible avec ces sous-là.

C'était quoi le pactole pour qu'on ait une idée ?

7 000 euros que j'ai mis de côté en 4 mois. Ces sous-là, je les ai finis vraiment en janvier-février 2014.

Et tu étais hébergé ? Tu vivais à...

Chez ma mère. Enfin, je lui paye un loyer, c'est 300 balles, mais bon, c'est pas comme si je louais un appart' quoi. C'est assez raisonnable. J'avais encore 1 000 balles qui dormaient en France, donc je les ai utilisés jusqu'à la fin de l'année et je me suis dit « bon ben là, la 3e année de licence, il va falloir que je bosse à côté ». Comme je suis un peu procrastineur parfois, je m'en suis occupé au dernier moment, c'est-à-dire mi-août avant la rentrée quoi, et l'idée c'était qu'il fallait un boulot avec des horaires flexibles pour les cours, et si possible aussi qui puisse être fait *at home* parce que j'ai pas le permis, et que j'ai pas envie de passer des heures dans les transports, ça me saoule. J'ai été sur LinkedIn, j'ai cherché un peu, j'ai trouvé ce truc-là…

Y'a des offres sur LinkedIn ? Je croyais que c'était juste une banque de CV...

Non, t'as aussi des entreprises qui mettent des offres d'emploi. Toutes les semaines, je reçois des offres.

T'avais déjà le CV en ligne sur ce réseau ?

J'ai refait mon CV, bien, et je l'ai remis à jour avec l'expérience Apple. Et j'ai trouvé la boîte SuccessTech. Donc je me suis inscrit sur leur site.

Alors là il me faut du détail. T'as tapé quoi comme mot pour trouver cette boîte ? Tu vois ? Comment t'en es arrivé à...

Je sais plus ce que j'ai tapé dans LinkedIn pour trouver ce truc-là, mais c'est sûr que j'ai dû taper « flexible », « *at home* », c'est les deux trucs. Ça a dû prendre des heures, hein, j'ai pas trouvé immédiatement. Avec ces *keywords*, j'ai trouvé deux ou trois trucs, j'ai postulé en premier à SuccessTech.

Alors, SuccessTech… ? Tu sais quoi de cette boîte ?

C'est une boîte qui vend des minutes de travail à des grands groupes – Microsoft, Google en premier lieu. En fait, si tu veux, je pense que ça fonctionne comme ça : t'as une boîte comme Google par exemple, ils ont besoin de milliers de gens pour faire un travail de fourmi, c'est vraiment un travail de fourmi. Je pense que,

de leur point de vue, engager vraiment des gens pour ça, c'est super contraignant, parce que c'est du travail hyperflexible, la quantité de travail disponible, elle varie vachement en fonction des jours et des semaines, donc avoir des gens en CDI ou en contrat permanent, je pense que c'est vraiment pas du tout adapté à mon avis. Et je pense que c'est pour ça qu'ils font appel à une boîte qui est spécialisée dans le fait de lever des armées et de vendre des minutes de travail et eux, c'est leur métier quoi. […] Et donc pour revenir à comment je les ai trouvés, j'ai trouvé l'offre, je me suis inscrit sur leur site, j'ai donné mon CV, ils sont revenus vers moi en me disant que mon CV correspondait et que je pourrais éventuellement être pris. Ce qui se passe à partir de là, c'est que t'as un examen à passer. En fait, tu reçois un PDF qui fait genre 50 pages, c'est l'information. Ils te disent : « Tu te démerdes, tu lis ça, c'est tout le taf expliqué étape par étape, tu lis, on te fait passer un examen, t'as compris, t'es pris. »

C'est le fameux guide qui a fuité sur Internet il y a quelque temps ? Les quality raters guidelines *qui ont fuité ?*

Ah ouais ? Je savais pas du tout. En fait, y'en a plein de *guidelines* (consignes). Parce que pour chaque type de tâche, t'en as une. Ça doit être le truc le plus général qui a fuité.

Et donc, ils t'envoient ces consignes...

Ils t'envoient ce truc-là, ils te disent « Bon voilà, l'exam c'est tel jour telle heure, vous vous connecterez, vous recevrez le fichier PDF, et vous aurez une semaine pour rendre l'examen. » C'est un truc en ligne, t'as d'abord un QCM pour voir si t'as vraiment compris les grandes lignes, et après t'as du taf, et en fait, je savais pas à ce moment-là, c'est exactement les mêmes tâches que tu fais quand tu bosses. Donc en fait, la formation c'est *crash and burn* : on te donne les documents explicatifs, t'as compris, t'as compris ; t'as pas compris, bah, basta quoi ! Donc en fait c'est *self-formation*, *self-service*, quoi. Voilà, moi j'ai fait ce truc-là, j'ai été pris. Tu reçois tes *login*, on te dit que tu peux commencer, et c'est tout en fait. C'est super expéditif.

Et la partie contractuelle, tu m'as dit qu'ils t'avaient demandé d'être auto-entrepreneur ? Comment ça marche ?

On te demande d'être employé indépendant. En France, je connais pas d'autres régimes qui permettent de faire ça, mais bon, moi j'ai pris auto-entrepreneur, hein. C'est génial, en dix minutes t'ouvres ton entreprise, ça prend deux clics, c'est vraiment bien. Donc on te demande ça et en fait ce qui se passe, c'est que t'es chronométré, t'as un logiciel qui te dit exactement à la seconde près combien t'as travaillé. Tu rentres à la fin de la semaine tes heures dans tes feuilles de temps, et à la fin du mois, ils t'envoient une facture les mecs, comme si c'est toi qui avais facturé à la boîte, même pas des heures de travail, mais des produits, sauf que les produits c'est... Le nom du produit, ça va être « heures de travail », et la quantité, ça va être 70 par exemple.

Et tu avais fait d'autres recherches de boulot en parallèle ?

Ben j'ai cherché un peu largement, et eux, c'est les premiers que j'ai trouvés. Et j'ai été pris, donc j'ai eu un beau coup de chance quoi.

T'avais pas fait d'autres candidatures ?

Non. J'en ai fait ultérieurement parce que j'essaye de travailler pour plusieurs à la fois, mais là, non, j'en ai pas.

T'as rencontré les gens de SuccessTech, ou tout a été fait à distance ?

Tout est fait à distance. Tu communiques par mail, y'a des vrais humains qui te répondent quand t'as des questions, mais tu les rencontres pas.

Et ils sont où ?

Ils sont en Irlande[8]. Pour des raisons fiscales.

Tu interagis en anglais ?

Tu interagis en anglais, toujours. Après, y'en a des Français quand même. Tu vois par exemple là, juste avant de venir, une fois tous les *x* mois, ils font une petite conférence, *conf call*, tu vois, ils te font une petite piqûre de rappel par rapport à certains trucs, et là, juste avant de venir, justement j'étais dans un truc comme ça. On est sur un truc, c'est pas Skype mais c'est similaire, et puis on papote entre nous. Et là, pour le coup, quand on fait ça, c'est un présentateur français pour les Français, un Anglais pour les Anglais. C'est le seul moment où tu interagis en français avec eux. […] Les factures par exemple [il me montre sa dernière facture], c'est SuccessTech qui produit ce truc-là, sauf qu'ils savent très bien ce qu'ils font. La facture, elle est comme si c'était moi. Il y a mon nom en haut à gauche, c'est comme si moi je leur avais fait une facture. Et ils se mettent en *invoice recipient*, ils se mettent en réceptionnaires de la facture alors que c'est eux qui la font quoi. Et dans le détail par exemple : *item* (le produit acheté), quantité, mais en fait la quantité elle correspond au nombre d'heures que j'ai faites. Et en fait, tu peux pas acheter 94 virgule quelque chose produit, c'est pas possible. Tu vois, c'est complètement du salariat déguisé en fait. Et du coup, sur ce que je touche, c'est moi qui paye derrière les charges sociales en fait.

Alors là t'as 0,6 h, c'est-à-dire 40 min de review task *?*

Ouais voilà, c'est ça. Les lignes, là, c'est différents types de tâches.

Et ils ont scindé ton travail en trois types de tâches, pour un mois ?

Oui, c'est un mois ça, c'est le mois de mars.

---

8. Dans un email envoyé le lendemain de l'entretien, Émilien a tenu à rectifier ce propos : « J'ai dit une bêtise concernant SuccessTech pendant l'entretien. L'entreprise n'est pas basée en Irlande, mais aux États-Unis. Elle a en revanche une présence en Irlande, et c'est bien d'Irlande que viennent les virements bancaires par lesquels elle me paye. »

Donc en fait t'as fait essentiellement du experimental task *?*

Ouais, ça c'est ce que je fais le plus ouais. *Side-by-side*, j'en fais beaucoup aussi. *Web result*, ça t'en fais assez peu. Mais en général, c'est *experimental* et *side-by-side* que tu fais le plus quoi. […] Et tu vois, quand j'ai fait l'auto-entreprise, je me disais que c'est vrai que c'est un peu du salariat déguisé en gros. Mais quand tu fais ton auto-entreprise, tu choisis tes statuts et donc t'as une liste du type d'activités que tu vas faire. Et dans la liste, ça m'a fait halluciner quand j'ai vu ça, y'a exactement le nom de ce taf-là sur le site de l'auto-entrepreneur français : « analyse de performance – assesseur internet indépendant ».

1.4. Rapport au travail : « C'est comme un panier de travail »

Donc là on a parlé du recrutement, du statut... Et la première fois que t'as bossé, comment ça s'est passé ?

Bah... Comment ça, je comprends pas la question.

Il fallait que tu te connectes sur un... ?

Oui, t'as un site, t'as tes *login* [il me montre]. T'as une plate-forme de travail, et t'as le travail qui tombe. C'est comme un panier de travail. Y'a du travail qui tombe dedans. Quand t'arrives, y'en a de dispo ou pas, et tu dois choper le travail quoi.

Un panier ? C'est curieux ça.

C'est ça.

Ah, et ton espace de travail, ça s'appelle le Rater Aid !

Ouais, c'est ça. Ça c'est un nom de domaine qui appartient à Google, si je dis pas de bêtises. [...]. Alors tu vois, là j'ai plusieurs trucs disponibles dans le panier. Ça, c'est la page d'accueil de l'espace de travail. T'as des tâches disponibles. *Experimental*, tu sais jamais ce que ça va être. Comme son nom l'indique, ils font des expériences quoi. Quand tu cliques sur *experimental*, tu ne sais pas à quoi t'attendre.

Et alors, est-ce qu'il y a des trucs que tu ne fais pas ? Les trucs qui apparaissent dans ton panier, si tu les fais pas ça fait quoi ?

Tu veux dire si par exemple il y en a que je fais jamais ?

Ouais.

Non, ça pose aucun souci.

Ils vont disparaître du panier ?

Non, d'autres gens vont les faire. C'est pas *mon* panier, c'est le panier de tout le monde en fait.

OK. Et alors, il n'y a que trois tâches ?

Non, là il y a trois *types* de tâches différentes. Imagine qu'il y a *experimental*, *result review* et puis un autre. Tu vois, c'est des rubriques. Quand je vais cliquer, ça va commencer, je vais faire une tâche, puis une deuxième, puis une troisième, et je peux rester trois heures si je veux.

1.5. Un travail micro-segmenté : « ça va de 30 secondes à un quart d'heure »

Et les tâches, c'est des micro-tâches ?

Une tâche, ça va de 30 secondes à un quart d'heure. Ça dépend vraiment de ce que tu fais. Le volume de travail disponible est super changeant, et le travail en lui-même est super différent. Des fois t'as des tâches qui durent littéralement 30 secondes, et les plus longues que je fais moi elles durent 13 min 50. Et tu fais des trucs à l'oreille, tu fais des trucs au clavier, etc.

Et le taux horaire est toujours le même ?

Oui. Toujours. T'es payé à l'heure en fait.

T'es payé à l'heure ? Alors qu'est-ce qui se passe si tu fais exprès d'être un peu lent ?

Ben justement. C'est ça qui est magnifique. C'est que pour chaque tâche que tu fais t'as un AET (*average expected time*), un temps moyen attendu pour cette tâche ; c'est le chrono quoi. Tu dois faire la tâche en ce temps-là maximum. Et ça, tu vois, c'est un logiciel qui ne m'a pas été fourni par SuccessTech, c'est un *add-on* pour le navigateur Chrome qui a été fait par des *raters*. Et les mecs ont fini par monter leur boîte et faire un abonnement payant à ce truc-là parce qu'il est tellement utile... T'es censé à la base te chronométrer toi-même, noter sur papier à la seconde près le travail que tu fais, et ensuite rentrer tes heures. C'est juste impossible. Et eux ils ont fait ce truc-là. Moi du coup, je paye l'abonnement 5 euros par mois y'a pas de problème. Et tu vois, ce truc-là, il tracke tout mon travail, combien de tâches je fais de chaque type, exactement, et du coup, c'est trop bien pour rentrer tes heures.

1.6. Chronométrage et productivité : « je suis à 108 % de vitesse »

Alors, par exemple aujourd'hui, tu as travaillé ?

Oui, j'ai bossé, mais juste 3 min dans le train en venant. Regarde, par exemple, mardi ressemble plus à une journée habituelle. J'ai fait 35 tâches de type *experimental*, j'ai fait 11 tâches *experimental mobile*, c'est-à-dire sur téléphone.

Sur ton téléphone ?

T'es censé le faire sur téléphone, mais moi ce que je fais, c'est que je me mets en mode développeur sur Chrome, je peux émuler un téléphone. Parce que sinon ça saoule quoi. Tu perds du temps, et mon téléphone est un peu merdique.

Donc là, en fait, t'étais sur ton ordi tout le temps ?

Oui. Donc là, tu vois, j'ai fait 11 tâches expérimentales qui sont des tâches chronométrées à 1 min 30. J'en ai fait 11. Au total, j'avais 16 min 30 maximum. Je les ai faites en 15 min. Donc je suis à 108 % de vitesse. De productivité, tu vois.

On n'a pas les heures, là ?

Non, mais tu vois quand tu as commencé et quand tu as terminé. Et après si tu veux regarder le détail, tu as tout le détail, c'est génial. Les mecs qui ont fait ça, ils ont tout compris. Je faisais pareil chez Apple, je trackais tous mes résultats, parce que, en *call center*, t'as mille stats pour chaque truc que tu fais. Et du coup, j'avais fait tout un *reporting*. Je m'étais dit que pour trouver du taf, ce serait bien que je puisse présenter de manière graphique et simple à comprendre la qualité de mon boulot. Donc en fait, j'avais – c'est un truc illégal –, j'avais aspiré tous les mails que j'avais reçus là-bas, j'avais... C'est super mal sécurisé chez Apple, tu peux avoir accès au serveur central, tu fais ce que tu veux. J'avais récupéré toutes les données qui me concernaient, et je m'étais dit qu'en rentrant j'allais les analyser, les miner, et faire un *reporting* complet de mon expérience d'employé chez Apple. Sauf que j'ai perdu cet ordinateur-là avec toutes les données de chez Apple, ce qui me fout trop les boules, parce que j'aurais vraiment aimé faire ce truc-là. Bon, et tu me demandais comment ils font si tu travailles doucement. Ben justement, c'est ça qui est magnifique. T'es payé à l'heure, et les chronos, une fois que t'es rôdé, tu les tiens super facilement. Quand t'as 10 min pour une tâche, tu la fais en 5 min disons. Donc après, t'as le choix : soit tu attends 5 min que la tâche se finisse, soit tu cliques pour passer à la tâche suivante. T'as aucune *incentive*, t'as aucun intérêt à passer à la tâche suivante, puisque t'es pas payé à la tâche, t'es payé à l'heure.

Ah, donc tout le monde est à 102 % alors ?

Les autres, je ne sais pas. Je connais personne, j'en sais rien. Mais ce qui est sûr c'est que tant que t'es au-dessus de 100 %, t'es tranquille. C'est assez simple une fois que t'es rodé. Et moi quand je fais une

tâche en 5 min alors qu'elle en demande 10, je vais me fumer une clope, j'attends 5 min, je suis tranquille, tu vois. T'as mieux sinon ; t'as des tâches de 10 min, tu les fais en 5 min, ça veut dire que t'as 5 min de surplus à la fin de la première tâche ; 10 min après la deuxième ; 15 min après la troisième. Donc tu travailles vite, t'accumules 20 min de surplus, et après tu laisses tourner 20 min sans rien faire. Parce que le chrono peut passer dans le négatif. C'est pas un problème. Ce qui compte, c'est ton chiffre global de productivité dans la journée. Si ton chiffre global est au-dessus de 100 %, il n'y a pas de problème. Donc, si tu veux, tu peux bosser rapidement une heure et accumuler trois heures de surplus et te barrer de chez toi quoi.

Mais tu laisses le bazar connecté.

Tu laisses connecté, tu laisses le chrono tourner, et tu reviens à la fin. Je pense que eux, ils sont à un niveau d'abstraction où ils voient les chiffres généraux de ce que tu fais, mais ils regardent pas dans le détail. Parce qu'ils ont des milliers de gens. Et ce qui est génial, c'est que quand tu commences, t'as le chrono qui se lance, et le Rater Aid m'envoie des sms pour me dire quand j'ai plus de temps, pour me dire quand il y a des nouvelles tâches qui tombent. Le logiciel, il est juste parfait. Il est développé par une entreprise qui s'appelle RaterWare.

Je découvre tout un monde là...

Ouais, ben moi aussi, quand j'ai commencé…

Tu connaissais quelqu'un qui avait fait ce boulot-là ?

Pas du tout. En revanche, je trouvais ça un peu trop beau pour être vrai, et aussi quand tu fais les examens d'entrée, c'est assez long, tu en as pour une bonne dizaine d'heures. Et l'examen, c'est du boulot, en fait. Moi au bout de deux ou trois heures, je me suis dit « Mais est-ce que c'est pas un canular super élaboré pour faire bosser des gens gratos ? ». Et donc j'ai été sur LinkedIn, j'ai fouillé pour trouver des gens qui avaient fait le même boulot, je les ai contactés sur Facebook après, et il y en a un seul qui m'a répondu : « Oui, j'ai travaillé 2 ans pour eux, ils payent en temps et en heure, c'est sérieux. » Donc ça m'a rassuré et j'ai continué. Et c'est une grosse boîte. Je crois qu'ils sont en bourse…

Qu'est-ce qui se passe si tu ne prends pas de tâche dans ton panier pendant plusieurs jours ? Est-ce que tu restes dans le circuit ?

Si tu ne travailles pas plus de trois jours d'affilée, ton compte est bloqué, mais ça veut pas dire que je travaille tous les jours hein. C'est qu'en fait, quand un compte est inactif plus de 3 jours, ils le mettent en pause, mais tu peux, à n'importe quel moment, leur envoyer un mail pour leur dire que tu seras absent de telle date à telle date sans avoir à leur donner de raison. En octobre par exemple, je suis parti au Portugal une semaine, je leur ai envoyé un mail la veille : « Bonjour, je serai inactif de telle date à telle date » ; ils m'ont répondu « Bonjour, très bien », et pas de souci.

C'est comme Pôle Emploi !

Ouais c'est ça, exactement. Sauf que tu gagnes des sous. Et par rapport aux horaires, je suis tenu contractuellement de faire 10 heures minimum par semaine, jusqu'à 20 heures à ma convenance. Sauf que depuis décembre, chaque semaine, on reçoit un mail qui dit : « Vous pouvez faire jusqu'à 30 heures cette semaine, il y a beaucoup de taf disponible. » Et donc ça fait depuis décembre que je fais 30 heures. Il faudrait que je fasse un tour d'année pour voir si par exemple l'été il y a moins de boulot, mais en tout cas moi tant qu'il y a 30 heures ça me va.

1.7. Une tâche devant l'enquêteur

Tu vois, le chrono il a démarré, là, je vois le temps qu'il me reste pour cette tâche-là, 8 min 50. Ici, c'est le temps total que j'ai travaillé aujourd'hui, et là, c'est le temps qu'il me reste pas pour cette tâche-là, mais pour un temps global que j'ai accumulé. Tu vois, c'est un peu plus que les min, parce que j'ai accumulé 1 min de temps supplémentaire en travaillant ce matin dans le train. Ce type de tâche-là, c'est du *quality rating*, on te donne une page internet, tu dois aller la checker, et tu dois la noter selon toute une grille de notation fournie par Google. Tu vois, là, la page à noter c'est sur le site de *Sud-Ouest*, le quotidien. C'est un article de news sur le crash de l'avion, la page est récente. Bon, *Sud-Ouest*, de toutes façons, je sais qu'il est bien... Normalement, t'es censé analyser pleins de trucs et les noter ensuite. *Sud-Ouest*, je l'ai fait des tas de fois, donc je sais ce que je vais mettre ; ça me fait gagner du temps. Tu vois, on te demande : « Est-ce qu'il y a du porn ? Est-ce que c'est une langue étrangère ? Est-ce que la page n'a pas chargé ? ». Bon, là, je ne coche rien. On te demande « Quel est le but réel de la page ? » Donc là, tu t'emmerdes pas, tu mets « *a news article about a plane crash* ». On te demande : « Est-ce que cette page est malicieuse, dangereuse, trompeuse, ou manque d'utilité ? » Donc je mets « Non ». « Est-ce que le but de la page est de faire de l'argent sans fournir quoi que ce soit à l'utilisateur ? » Je mets « Non ». Et « Est-ce que c'est *ymyl* ? » Ça veut dire : « *Your money your life*. » C'est une catégorie qu'on t'explique dans les *guidelines*. C'est toutes les pages qui sont susceptibles d'impacter la vie future de l'utilisateur d'une manière qui pourrait avoir des conséquences légales graves : information médicale, information légale, transaction financière et informations sensibles. Tout ça c'est des sites où tu dois placer la barre au plus haut niveau de confiance. Là je mets « Non ».

Si c'est un site avec un formulaire à remplir, c'est ymyl *?*

Par exemple, si c'est un site où tu fais des achats, ça va être « *ymyl shopping* ». Pareil pour un site d'infos médicales. Et si le site n'est pas référencé ou si c'est pas un site sérieux, il va se prendre une claque. Mais pour *Sud-Ouest*, c'est pas le cas. Et ensuite tu as la grille à remplir, je te le fais très rapidement. On te demande « Qualité du contenu principal ».

Tu le lis pas ?

Non, là je sais que c'est court, c'est bien écrit, je vais mettre « medium ». Parce que c'est certainement correct, mais c'est court, c'est juste une dépêche quoi. Ç'aurait été un article de trois pages, j'aurais

regardé vite fait si c'est bien écrit. Mais tu vois, la source, je sais que c'est un truc plutôt sérieux donc voilà. On te demande ensuite « Expertise, autoritativité et confiance de la source », « Réputation du site web ».

Alors typiquement, pour Sud-Ouest*, tu mets quoi ?*

Là je vais pas m'emmerder, je vais mettre « *well-known website* », du coup j'ai pas à faire de recherches. J'en sais rien, je sais de quel bord ils sont, je m'en fous, mais c'est pas ce qu'on me demande. Je pense que c'est une maison d'édition sérieuse. Ils demandent ensuite « Informations sur le site web » : en gros est-ce qu'on trouve facilement les informations légales, le numéro SIRET de l'entreprise, qui est derrière, etc. Bon, là je sais que oui parce que j'ai déjà vérifié. « Est-ce qu'il y a du contenu supplémentaire ? » Supplémentaire, c'est quand le site te fait des propositions ciblées. Enfin, « Design de la page », c'est est-ce qu'elle est bien entretenue.

Et là, une fois que t'as fini, tu fais quoi ?

Tu donnes une note globale sur cette échelle, que tu justifies, et ensuite tu fais *submit* pour passer à la tâche suivante, ou *submit and stop* si tu veux arrêter de bosser. Bon, là, c'est un cas classique, une page d'un site connu, donc tout va bien. Et donc là, je vais laisser tourner 3 min, parce que je suis payé.

Et les autres tâches ?

Il y a un document d'entraînement par type de tâche, je te donnerai tout ça. Le truc que tu fais le plus, c'est le SxS, le *side-by-side*. Tu as une requête qu'un utilisateur a faite sur Google, et t'as deux listes de résultats de Google.

Produites avec deux algorithmes différents ?

Je sais pas exactement ce qu'il y a derrière, mais ça doit être ça, ou alors le même avec des *seeds* différents. En tout cas, t'as deux *outputs* différents, et on te demande de noter chaque résultat un par un, des deux côtés, et ensuite de donner une note globale à chacun des côtés, et de dire globalement lequel te paraît le mieux et pour quelle raison. T'as toute une série de critères des *guidelines* à prendre en compte. Un autre truc que tu fais souvent aussi, c'est juger juste un site, comme ce que je viens de faire avec *Sud-Ouest*. Il y a aussi le *safe search*, c'est du nettoyage sur YouTube : virer les trucs violents, s'assurer que les enfants peuvent pas trouver des saloperies, pleins de trucs comme ça…

1.8. Des pénibilités consenties

Et alors, t'es exposé à des images parfois un peu choquantes ?

Pas sur YouTube, parce que YouTube, ils sont assez forts pour éliminer automatiquement tout ce qui est porno, gore, etc. Mais après il y a des trucs légitimes. Par exemple, sur YouTube, une scène de film où il y a pleins de gros mots, ils vont pas virer ça, c'est légitime. Par contre, quand il y a le *safe search*, le contrôle parental qui est activé, il faut pas qu'un enfant tombe là-dessus. Donc moi ça ne me pose pas de problème d'être exposé à ça. En revanche, tu travailles pas mal sur le porno aussi... Parfois, pendant une semaine, je vais faire le même type de tâche.

C'est eux qui décident ?

Tu te connectes, tu vois ce qu'il y a de disponible, tu peux pas savoir à l'avance. Je ne sais pas quelle est leur manière de gérer ça, mais en tout cas, au bout de quelques mois tu te rends compte qu'il y a vraiment des périodes où tu vas plus faire un taf qu'un autre. Pour revenir à la question d'être exposé à des trucs choquants, franchement, je vais te dire : quand tu fais une semaine de porn, c'est relou quoi.

C'est quoi une semaine de porn ? C'est des sites à évaluer ?

Tu fais que de la notation de sites pornos, toute la semaine, à la fin t'en as marre.

Tu dois faire des drôles de rêves...

Y'a des moments où tu peux plus, tu vois, t'es... Et puis alors, tu vois des trucs... En fait, tu as aussi un truc où tu dois signaler tout ce qui est pornographie infantile et bestialité, etc. Bon, là, pareil, tu vois des trucs, euh.... Bon, OK... Et sinon, t'as aussi les *image boards*, c'est des espèces de forums, mais un peu particuliers, qui sont apparus sur la toile ces dix dernières années. En fait, c'est des forums où les gens parlent, mais à chaque post, t'es obligé de mettre une image. Et ce qui se passe, c'est que les *threads*, les [fils de] discussions, c'est jamais des discussions par rapport à quelque chose en particulier, c'est des discussions par rapport à un type d'image. Et donc ça donne lieu à des trucs super comme à des trucs vraiment horribles aussi. C'est des sites assez obscurs en général, et donc du coup t'as des gens de tous types qui se retrouvent là-dessus : des gens qui veulent partager des œuvres d'art, mais aussi des gens qui veulent partager du *cheese pizza* comme ils disent, CP, c'est le nom de code pour *child pornography*. Voilà. Tu peux sortir ça en soirée mondaine.

Du coup Google les épingle ?

Ouais, ces trucs-là, tu les signales. Il y a un site de ce type-là qui a été fermé il y a 2 mois, et je sais que je l'avais noté des tas de fois, je lui avais mis des claques parce que c'était incroyable, il y avait des trucs hyper *borderline*, même carrément pornographie infantile.

Quand ton panier te propose des tâches, il te prévient si ça va être du contenu porno ou pas ? Parce que si t'en as marre du porno...

Quand tu vas cliquer sur *experimental*, t'arrives sur la page, et là il y a écrit en rouge : « Attention cette tâche contient du porno, donc si vous voulez pas faire de porno, ben vous la faites mais vous arrêtez après. » Si tu te lances, t'es obligé de faire celle que tu viens de lancer. Cela dit, t'as des cas où t'as le droit de *release*, de lâcher la tâche si elle n'est pas dans la bonne langue, s'il y a un problème technique, etc. Mais je ne crois pas que le fait que ça soit du porno te donne le droit de lâcher la tâche.

Tu dis que t'en as fait pendant un mois. Ça veut dire que ça te dérangeait pas trop ?

Non, mais c'est que j'ai besoin de thunes et je ne peux pas dire non au taf quoi.

Il n'y avait rien d'autre ?

Ben par exemple, la semaine dernière, j'en ai fait beaucoup, et en fait c'est pas qu'il n'y avait rien d'autre mais les autres trucs c'était plutôt des tâches de 30 secondes, et c'est pas des tâches où tu peux travailler 5 min, et t'arrêter 5 min, etc. Et vu que je fais d'autres trucs à côté aussi, j'aime bien avoir des tâches qui sont longues, où le chrono est facile à tenir, parce que ça me permet de faire d'autres choses à côté, de faire ma journée aussi, tu vois. Et il y a un autre truc par rapport au porn, c'est que dans les options, il y a un item « *I am willing to participate in tasks with adult content* », tu peux le décocher. Dans ce cas, tu n'as pas de porno, mais tu vas perdre du travail du coup.

Est-ce que si tu décoches cette option, il y a un risque que par moment tu n'aies rien ?

Si à un moment donné, il n'y a que des tâches *porn-related* qui sont disponibles, alors oui, t'auras rien. Mais en général, y'a du taf disponible, franchement. Ça arrive, mais c'est rare, que tu te connectes, et qu'il y ait écrit en rouge « Il n'y a pas de boulot ». Mais ça m'est arrivé 3 ou 4 fois depuis que je bosse pour eux, ça a duré un jour.

1.9. Une combine à 3 200 euros par mois

C'est quoi tes préférences entre les différentes tâches qu'ils te proposent ?

Ça dépend d'où je suis et de ce que je suis en train de faire. Quand je suis chez moi et que j'ai plusieurs heures devant moi, ce que je préfère faire, c'est le *page rating*, parce que c'est simple, ça va vite, et les chronos sont faciles à tenir. Ou du YouTube, parce que c'est facile de tenir les chronos aussi. En tout cas, quand je suis chez moi, je préfère les tâches les plus longues possibles avec les chronos les plus faciles. C'est là que tu peux vraiment faire des heures de travail, faire une grosse journée. Et vaut mieux pas faire 4 heures d'affilée des tâches de 30 secondes, tu te tires une balle quoi. En revanche, quand t'es avec des amis... C'est pas évident de faire 30 heures de taf en plus des cours à la fac. Du coup, souvent, je vais chez un pote, on se fait un Mc Do,

moi j'ai l'ordi à côté, et là je bosse. Et dans ce scénario-là, ce que je préférerais, c'est les tâches où tu réfléchis le moins. Parce que t'as des tâches où tu peux être totalement absent intellectuellement. Tu regardes, tu fais deux clics, boum, c'est fini. T'as 1 min 30 pour les faire, et…

Et c'est quoi ? Tu bouffes le McDo, tu discutes avec tes potes, et en même temps tu bosses ?

Bah oui, tu bosses et puis, toutes les 1 min 30, tu fais deux clics en 2 secondes, et t'as 1 min 30 devant toi.

Et ils disent quoi tes potes ?

Ils se sont habitués quoi.

Et au début, ils ont dit quoi ?

Euh, au début, je sais pas ce qu'ils ont dit. Je me souviens pas.

Ils t'ont charrié, c'est sûr !

Ouais, je sais plus, franchement, je saurais pas te dire.

Ou alors ils ont dit : « Vas-y, montre-moi » ?

Ben il y en a qui se sont intéressés, notamment Olivier, que je viens de faire rentrer en l'aidant à passer les exams. Parce qu'en fait, vu que les chronos sont faciles à tenir, moi depuis des mois, je me dis : « Putain, il me faut un deuxième compte chez eux. » Parce que, imagine, j'ai 10 minutes pour faire un tâche. Si j'ai deux fenêtres ouvertes, je vais travailler le même temps, mais être payé le double !

Ouais, mais tu vas te faire griller, non ? T'as la même adresse IP.

Ouais, mais j'utilise un proxy donc il n'y a pas de problème. Ça permet de te faire passer pour une autre adresse IP. Et là je viens de faire rentrer Olivier chez SuccessTech, il va commencer à travailler cette semaine, sauf que le deal c'est que lui, il ne fera jamais plus de 10 heures par semaine. Moi je ferai les heures restantes, et on se partagera les sous au prorata de qui a travaillé combien d'heures. Et c'est super intéressant parce que ça veut dire que je vais pas bosser deux fois plus pour gagner deux fois plus ; je vais doubler mon salaire horaire. Et bon, 30 heures, avec les cours en plus, c'est un peu compliqué, mais quand j'aurai plus les cours, 30 heures de boulot par semaine c'est pas énorme. C'est même pas une semaine de temps complet. Tu fais tes horaires comme tu veux, t'es vraiment flexible, tu fais ce que tu veux. Je veux dire, c'est pas méchant comme boulot. Là, je suis payé 1600. Être payé 3 200 pour ce taf-là, moi je signe direct ! Ça veut dire que tu peux aller à Paris en terrasse, passer deux heures au soleil à bosser, ça veut dire que tu peux te balader en France partout où tu veux, avec un proxy, ça veut dire que tu peux aller à l'étranger si tu veux. Bref, tu peux te faire un tour d'Europe ! Avec mon pote, on s'est dit : « On

s'achète un Van. » Y'en a un qui conduit, l'autre qui bosse, on alterne toutes les deux heures, et t'es payé en chemin. Tu vois, moi je m'étais dit : « Il faut de la flexibilité horaire et spatiale pour les cours à la fac. » Mais j'avais pas vu aussi loin en fait. Bon, c'est un travail qui est précaire, parce que j'ai pas de sécurité d'emploi. Ils m'envoient un mail demain pour me dire que c'est fini et c'est fini. Mais tant que ça fonctionne, franchement, ça a des super bons côtés. T'as aucune obligation.

T'as un plafond de revenus en tant qu'auto-entrepreneur, non ? C'est combien les revenus maximum ?

C'est 32 000 euros par an je crois. Mais bon, avant que je les fasse... Faudrait que je fasse 2 500 euros par mois ; j'y suis pas encore.

1.10. « Des vraies données »

Est-ce qu'il y en a d'autres à la fac à qui t'as donné le tuyau et qui se sont mis à faire ce taf aussi ?

Non, les gens ils me voient bosser en cours, on m'a déjà posé la question « Qu'est-ce que tu fous ? », mais personne m'a demandé de les faire rentrer. Tu sais, les gens ils voient ça, ils se disent... Déjà dans ma classe, je crois que les gens me voient un peu bizarrement. Tu sais, je suis le gars qui est sur son ordi, qui écoute pas mais qui va poser des questions d'un coup qui vont être plus pertinentes que les autres, et les gens ils se disent « Mais c'est quoi ce mec ? » Et je sais que parfois on regarde ce que je fais, mais ils doivent se dire que c'est un truc d'informatique.

Un truc de geek ?

Ouais, c'est ça. En fait, c'est pas un truc d'informatique. Ce que tu fais, c'est noter des pages. Ça ne demande pas de connaissances en codage, c'est pas de la programmation, c'est pas de l'informatique. Mais c'est sur ordinateur, c'est Google, donc les gens doivent se dire... Mais non, on ne m'a jamais demandé. Et donc globalement je t'ai à peu près présenté le truc. Et juste pour te faire une petite liste des types de tâches qui existent, t'as les *side-by-side*, t'as une requête de l'utilisateur, tu notes deux listes de résultats différentes. T'as aussi le *page result* qu'on a fait avec *Sud-Ouest*. Ces deux-là, c'est les principaux. Ensuite, t'as *book search*, c'est pour les gens qui font des recherches sur Google Books. Ils font une recherche et on vérifie si c'est le bon livre, la bonne page, est-ce que c'était utile ou pas, est-ce que le résultat est pertinent. Il y a aussi *duplicate images*, donc là tu fais de la chasse au doublon dans les images dans Google Images. Donc on te présente un mur d'images et c'est comme les jeux de mémoire, comme un Mémory : tu sélectionnes les doublons et voilà, c'est tout. C'est assez facile d'ailleurs.

Et ça, c'est des petites tâches qui se font très très vite ?

Non, ça c'est un quart d'heure. Mais c'est parce que t'as un mur d'images énorme. Tu vois une requête qu'un mec a fait, t'as un mur d'images énorme, tu dois sélectionner les doublons pour qu'il ne les présente pas en double. Tu dois voir dans le doublon, lequel est de meilleure qualité. Ça

prend du temps quand même. Il y a aussi le « Google Now », ça c'est ce qu'ils appellent les *device actions*. Attention, là on rentre dans le genre de truc qui fait flipper les gens quand je leur dis. C'est quand tu notes la pertinence d'une action que le téléphone a faite quand une personne lui a demandé, à son téléphone, de faire quelque chose. Typiquement : « Envoie un message à Sarah, dis-lui que j'arrive dans 15 minutes. »

Il peut faire ça, le téléphone ?

Ouais. Donc moi je passe derrière, j'entends le mec dire ça, je vois où il était, à quelle heure il l'a fait, et qui reçoit le message.

C'est des vraies données ?

C'est des vraies données. Ah, tu l'entends, le mec, tu l'entends parler ! Ça, tu vois, c'est limite [il s'esclaffe], c'est un peu flippant. C'est anonymisé, mais tu vois l'heure exacte, l'emplacement à la rue près sur Google Maps, et puis tu l'entends la personne. Mais en même temps, bon, il a cliqué sur « J'accepte les conditions d'utilisation d'Android », voilà, quoi. S'il dit à son téléphone : « Emmène-moi rue d'Ulm », et bien moi il faut que je vérifie que le téléphone a bien lancé la requête sur le GPS, qu'il a bien calculé le bon itinéraire, etc., des vérifications comme ça.

C'est des trucs de reconnaissance vocale ?

Oui, le Google Now, c'est toute la partie reconnaissance vocale. T'as aussi « synthèse vocale ». Là, tu vois du texte et t'as Google qui te lit le texte, et tu dois juger des échantillons et dire lequel paraît le plus humain, pourquoi, etc. Bref, tu juges de la synthèse vocale. Tu fais aussi de la traduction et du jugement de traduction. On te présente un texte en français, tu le traduis en anglais, ou l'inverse. Et pour le jugement de traduction, on te propose deux traductions différentes d'un même texte, et tu dis laquelle est la mieux, et pourquoi. Et en détails, hein. T'as toute une grille avec les rubriques « erreur de choix de mot », « erreur d'ordre des mots », « erreur morphologique », « erreur de conjugaison », « erreur de grammaire », « erreur de ponctuation », etc. Donc parfois ça peut prendre un peu de temps parce que, sur une seule phrase, tu peux en avoir des dizaines. C'est comme pour ce qui est vocal : tu peux juger la reconnaissance et la synthèse, et là c'est pareil, tu peux fournir ton expertise en traduction comme tu peux juger de la qualité de la traduction que Google Trad fait. Il y a aussi *Mobile needs*, c'est quand les gens font des recherches locales, quand ils demandent des choses près de chez eux. Tu vas vérifier en fonction de la position de l'utilisateur la requête qu'il a faite, les résultats qui ont été donnés, et tu dois dire si c'est pertinent ou pas. C'est toujours ça, c'est toujours des questions de pertinence. Ensuite, il y a *Send to device* : tu notes des pages, mais sur téléphone. Il y a YouTube aussi : tu fais du rangement sur YouTube, tu vires les contenus illicites ou dangereux. Et il y a d'autres petits trucs qui apparaissent parfois dans le panier mais c'est des trucs expérimentaux. Tu vas les voir une semaine, ils vont disparaître.

1.11. Multiplier les employeurs

Donc en fait, des gens de Google, t'en a pas rencontré ?

Non, jamais.

Ton seul interlocuteur, c'est SuccessTech ?

Ouais.

Tu m'as dit que tu avais cherché à trouver un équivalent dans une autre boîte qui fait plus ou moins le même truc.

Oui. J'ai postulé à une boîte qui s'appelle DataGate. Eux, pareil, ils font du *crowdsourcing*. Ils lèvent des armées de gens et ils vendent de la minute de travail. J'allais être pris, et eux, contrairement à SuccessTech, ils font un entretien par Skype avec une vraie personne en face de toi, et pendant cet entretien, ils m'ont demandé si je travaillais pour quelqu'un d'autre dans le même domaine. Je savais parfaitement qu'il fallait que je dise non puisque j'avais déjà menti sur mon CV pour ne pas leur dire. Et je sais pas ce qui m'a pris, j'ai dit : « Oui, bien sûr, je travaille aussi pour SuccessTech. » Et je me suis dit « Pourquoi tu viens de dire ça ? » En fait, c'était un gros lapsus révélateur, et du coup j'ai pas été pris. Ils m'ont suggéré de me séparer de SuccessTech pour me prendre, et je le ferai pas parce que je suis avec SuccessTech, donc voilà. Mais je sais pas pourquoi je leur ai dit, j'ai vraiment été con.

Tu sais pourquoi ils sont exclusifs, pourquoi ils veulent que tu quittes SuccessTech ?

Je ne sais pas pourquoi, mais il y en a d'autres aussi que j'ai essayés, et tous veulent que tu ne bosses pas pour quelqu'un d'autre. T'es auto-entrepreneur, tu fournis un service à des entreprises, ils ne peuvent pas t'en empêcher en principe.

Cet empêchement, c'est peut-être interdit ?

Ah ben oui, clairement. En plus, dans le régime d'auto-entrepreneur, t'as pas le droit d'avoir un seul client plus d'un an, justement pour éviter le salariat déguisé.

Alors comment tu vas t'y prendre ?

Ben je vais faire une facture de 500 balles à un pote pour un service informatique, voilà quoi.

Et les autres boîtes que tu as essayées ?

J'ai essayé MaxSoft aussi.

Où as-tu trouvé ces noms d'entreprise ?

Sur Google, j'ai tapé *crowdsource job*.

Ah, c'est ça le mot-clé...

Non, pardon, en fait, quand j'ai rempli les formulaires de DataGate, au moment où ils m'ont demandé si j'avais bossé pour d'autres boîtes, il y a la liste complète de toutes les boîtes qui font ça. Et donc MaxSoft, ils m'ont pris. Leur client, c'est Microsoft par contre, pas Google. Et il y a très, très peu de travail disponible. Je fais 10 balles par semaine, donc en fait ça me sert à rien.

Ah, donc t'as quand même deux trucs en parallèle ?

Ouais. Mais y'a rien chez MaxSoft. Et eux, en plus c'est payé à la tâche, pas à l'heure. Donc c'est beaucoup moins avantageux. Du coup, j'y vais plus trop.

1.12. Une tentative de distanciation : « je suis le dernier ouvrier de la chaîne »

Bon, et qu'est-ce que tu penses de ce travail, de ce que tu fais ?

Tu veux dire l'utilité de ce boulot, mon expérience ?

Les deux. Qu'est-ce que tu penses de tout ça quoi.

Bah. En tant que source de revenus, c'est clairement pas *secure*, t'as aucune sécurité de l'emploi. Le jour où Google aura développé une nouvelle technologie automatique, on va tous se faire virer. Même, si tu fais une connerie, ils te virent, ils s'en foutent quoi. Donc c'est pas un job pour tout le monde. C'est pas pour les gens qui ont des obligations. Mais c'est vrai que tu peux faire des thunes quand même. Parce que 1600 euros pour trente heures, c'est bien. Bon, en brut, hein, donc ça fait 1400 net. Et t'as le choix sur tes horaires, t'as pas le client en face de toi, etc.

Ça doit faire rêver des tas de gens.

C'est clair.

Surtout qu'il y a un côté presque...

C'est un jeu un peu. Ça demande pas une concentration extrême. Mais je tape super vite sur le clavier donc c'est plus facile pour moi, surtout que certains chronos sont pas évidents à tenir quand tu tapes normalement. Donc moi dans mon cas, ça me plaît beaucoup. Parce que c'est simple. J'ai peut-être moyen de faire le double horaire si j'arrive à mettre en place ce qu'il faut. Là ça sera juste parfait.

Tu vas gagner plus d'argent que moi.

Déjà rien que là, je gagne plus que ma mère, alors que je fous rien et qu'elle, elle se casse le cul, je me dis que c'est magnifique.

Elle fait quoi comme boulot ?

Elle est agent immobilière. Avec un deuxième compte, je vais être payé 3 200 balles pour 30 heures de taf par semaine, où je veux, quand je veux, comme je veux. C'est déjà mieux que 90 % des gens salariés dans le pays quoi.

Et tu penses quoi de la signification de ton travail, du sens de ton travail (je cherche pas à te juger ou quoi que ce soit) ?

Le sens, ben... c'est du complément d'intelligence humaine. T'as les algorithmes qui font le travail de gros. Nous ce qu'on fait, c'est du *fine-tuning*. Du réglage de marge. Comment dire... Une manière intelligente de présenter où se trouve ce boulot sur la frise chronologique de l'avancement de l'intelligence artificielle et de la société de l'information... En fait, un algorithme, profondément, c'est une suite d'instructions. Une suite d'instructions linéaires, c'est une chose. Après, tu peux faire une suite d'instructions avec des conditions. Là, t'avances un peu. Après, tu peux éventuellement faire une suite d'instructions qui, avec certaines conditions, est capable de créer de nouvelles instructions à partir de morceaux d'instructions que toi t'as déjà fournies. Là tu commences à avoir des systèmes qui sont malins, capables de s'adapter. Mais c'est toujours un système qui ne sait absolument rien faire de plus que ce que tu vas le préparer à faire.

D'accord, mais je cherche le lien avec le sens de ton travail.

Justement, tu peux faire un algorithme qui recombine des choses avec des tas de conditions. Mais ça restera toujours qu'une suite de choses qui auront été préparées à l'avance par un être humain. Quand tu veux vraiment faire de l'intelligence artificielle, là t'es obligé de prendre les choses différemment. Mon taf, il existe parce que, aujourd'hui, on n'est pas capables de faire des machines aptes à comprendre le contexte des choses, d'aller plus loin que les instructions qui leur ont été données, et donc on est obligés de faire appel à de l'intelligence humaine pour faire le boulot que je fais. Si tu me demandes ce que c'est mon boulot et ce que j'en pense, je dirais que c'est une des étapes nécessaires dans le chemin vers l'intelligence artificielle. Tu vois, aujourd'hui, les chaînes de montage de voitures automatisées ? Avant, c'était des gens qui le faisaient, ben je suis un peu cet exécutant-là sur la chaîne, qui un jour sera remplacé par un robot. Sauf que ça sera un robot d'intelligence artificielle. Et ce que je fais moi d'ailleurs, c'est du travail à la chaîne. C'est là qu'ils ont été intelligents chez Google. Quand ils ont fait leur moteur de recherche, ils se sont pas dit : « On va faire un annuaire de sites rempli à la main. » Avant ça existait, les annuaires de sites avec des gens qui remplissaient les rubriques, qui notaient les pages, etc. C'était avant le PageRank, qui fait ça automatiquement. Avec l'algorithme, c'est du passage à la chaîne. Moi, je suis le dernier ouvrier de la chaîne, qui est encore là parce qu'il y a encore un dernier truc qu'on ne sait pas faire faire à une machine. Je pense que c'est la vue globale la plus reculée que je peux faire sur le boulot que je fais.

Et donc, SuccessTech, ils savent beaucoup de choses sur ton travail, ils savent combien de temps tu consacres au boulot, comment tu fais les ratings*, etc. Mais toi, tu sais très peu de choses à la fois sur SuccessTech et sur Google. Tu sais pas comment il marche, leur fameux PageRank...*

Ouais. Le truc, c'est que parfois, t'as des requêtes faites par les utilisateurs que tu dois traiter, que tu n'es plus censé revoir par la suite. Mais t'en as qui reviennent régulièrement. À force de bosser, tu commences à avoir des schémas de répétition, des régularités, qui te font te dire que si ces trucs-là reviennent régulièrement, c'est que l'algorithme a du mal avec. Donc déjà tu vois ce qui pose le plus problème aux algorithmes. À mon petit niveau, j'ai pas une idée globale. Je fais aussi des notations d'articles de presse. Lequel est le meilleur, en fonction de la qualité de l'écriture, de l'autoritativité, etc. Autant juger deux listes de résultats, c'est une chose, mais juger des articles de presse, c'est différent. Si y'en a un auquel je mets une note de merde et un autre auquel je mets une bonne note, quand tu vas faire une requête sur Google, celui à qui j'ai mis une bonne note il va arriver en haut, et l'autre en bas. Ça a un impact tu vois, ce que je fais. Et je ne trouve pas ça normal en fait. Google, c'est fou le centre de pouvoir que c'est. C'est un truc de dingue. Je ne sais pas si tout le monde réfléchit à ça.

Moi j'y réfléchis pas mal !

Pour le meilleur comme pour le pire, hein. Mais on est tellement habitués à simplement aller sur Google pour chercher une info... Chez Google, ils font un truc qui s'appelle le *bubbling*, le « bullage » si tu veux. Ils analysent ton comportement sur les moteurs de recherche, les sites que tu visites, etc. Et tu vas sur deux ordinateurs différents, tu tapes la même requête, et tu ne vas pas avoir les mêmes résultats. C'est bien, parce que moi j'utilise Google souvent. Mais du coup, tu t'enfermes petit à petit dans un jardin fermé. Au lieu de t'ouvrir des horizons énormes, en fait Google te connaît de mieux en mieux pour resserrer l'horizon au plus proche de ce que tu fais. C'est pratique, mais en un sens, si tu le sais pas... Faudrait que les gens le sachent en fait.

1.13. La mise à pied et le bonus

Si tu as encore du temps, je voulais parler des formes de contrôle de ton travail, au-delà du chronométrage.

Ce qui se passe, c'est qu'il faut évidemment qu'ils aient une idée de ton travail. Mais ils vont pas faire passer quelqu'un derrière toi, sinon ça fait une personne en trop qui bosse. Ce qu'ils font, c'est qu'à la fin de chaque mois, t'as un gars qui passe derrière toi, qui prend une dizaine de tâches que t'as faites, au hasard, qui les note, selon des critères exacts des dossiers d'entraînement, et qui te donne une note globale. Et il faut que tu sois au-dessus de 85 % de correct. Si t'es en-dessous mais que tu dépasses 65 %, on te donne des entraînements à faire avant que tu puisses re-bosser. Donc tu vas devoir faire 5 heures d'entraînement.

C'est comme le code de la route alors !

Ouais, c'est ça. Si tu le fais deux mois d'affilée, t'es mis à pied, c'est ce qu'ils disent. Moi ça m'est jamais arrivé. Ils disent « *your account will be placed under review* ». Ça veut dire qu'ils vont passer du temps à regarder tout ce que t'as fait. Je pense qu'ils ont conscience que tu fais des milliers de tâches par mois. Si c'est pas de bol, s'ils tirent un échantillon pourri de ton travail, ben ils vont vérifier plus en détail. Les deux premiers mois, ils sont plus indulgents, il n'y a pas de mise à pied. Parce qu'au début, tu fais n'importe quoi, franchement. C'est difficile de bien intégrer tous les critères qui te sont demandés et de savoir les appliquer à chaque fois. Quand le mec te demande des trucs simples (une date, un événement, un site web), ça va. Mais t'as des gens, sur Google, qui te sortent des trucs... Encore aujourd'hui, parfois je vois une requête, et je me dis « Qu'est-ce que je vais faire ? » Regarde, ce matin j'ai eu ça : « recherche sur la céramique de fouille de Jordanie du VII^e^ siècle au XII^e^ siècle après J.-C. ». Voilà. Le mec, il tape ça.

C'est incroyable.

Donc la requête, elle veut rien dire quoi. En fait, non : elle est hyperprécise. Mais les trois quarts des résultats vont être à côté de la plaque. Et il faut que je trouve un truc dans la liste de résultats qui ait quand même un rapport avec sa requête, qui donne un résultat pertinent. Et ça encore, c'est gentil. Mais t'as des gens qui tapent n'importe quoi. Et là tu sais pas quoi faire. T'as aussi des gens qui comprennent pas comment Google fonctionne. Ils vont taper : « Bonjour Google, s'il te plaît : trouve moi l'adresse de ma grand-mère. »

Et vous avez un forum de raters*, où des blagues s'échangent et tout ?*

Ouais, mais j'y suis jamais allé. On a un portail communautaire, mais j'y ai jamais foutu les pieds. [Il reprend son ordinateur, me montre le portail communautaire, puis remarque des statistiques sur le travail disponible.] Tu vois là, il y a 10 110 *experimental* disponibles, sachant qu'une *experimental*, ça peut aller de 30 secondes à 10 minutes. Et il y a 1 608 *side-by-side* disponibles. Une *side-by-side*, c'est 9 min pile. Là y'a 240 heures de travail disponibles. Et ça, c'est 13 % du total disponible. Donc, quand je te dis qu'il y a du boulot de dispo, il y en a. [...] Au fait, quand je te disais qu'il n'y avait pas d'incitation à travailler vite, mais que tu avais plutôt intérêt à laisser tourner les tâches jusqu'au *average expected time*, il faut que j'ajoute une précision. En fait, y'en a quand même une incitation. Tous les mois, il y a un bonus de 500 euros pour le meilleur du pays. Donc tu reçois un mail tous les mois qui dit qu'ils lâchent 500 balles à un mec du pays supposé le meilleur. Je ne sais pas quels sont les critères, je suppose que la productivité joue un rôle.

S'il n'y en a qu'un par pays, c'est pas lourd...

Non, c'est clair, mais si tu me dis qu'on n'est que 10, je vais peut-être me mettre à bosser mieux pour gagner 500 euros par mois en plus.

Revenons à la question de la liberté. Tu penses à quoi pour plus tard, t'as des idées sur ton avenir professionnel ?

Ce truc-là, j'ai vraiment besoin que ça marche jusqu'à septembre. Mais là, c'est top secret, je t'en parle pas. J'ai trop souvent parlé avant d'agir dans ma vie, et là, non, je veux pas. J'ai deux gros projets en incubation dans ma tête depuis un an / un an et demi, d'ordre informatique. Si ça marche, tant mieux. Si ça marche pas, je sais pas. Mon autre plan, c'est de faire un master en RH et de rentrer chez Microsoft en alternance. Mais au final, j'en ai marre des études, ça va pas assez vite, et ça me rend fou. Ça va tellement doucement que ça me frustre. Les gens, ils sont trop lents, pardon, mais... Et il y a aussi le fait que le salariat, ça me terrorise. J'ai vraiment pas envie d'asseoir mon cul sur une chaise, de me dire que j'ai cinquante ans de salariat devant moi... Donc je vais essayer d'être entrepreneur, de faire des projets.

Ton père s'en est sorti après la faillite de sa boîte ?

C'est compliqué. Il a perdu sa boîte. Maintenant il va mieux. Et après il est retombé sur un gars qui est à la tête d'un gros business de télécom, il l'a retrouvé par hasard... Peut-être. Enfin, il est franc-maçon mon père, à mon avis, ça doit être ça. Et il l'a fait rentrer dans sa boîte.

1.14. « Si c'est que ça, c'est pas méchant »

Et sinon, dans ton entourage, à part ton père, tu as d'autres contacts geeks ?

Quelques-uns chez Apple.

T'es assez isolé en fait ?

Assez, c'est vrai.

Parce que tu développes quand même une passion, des réflexions poussées sur la sécurité informatique...

L'informatique pour l'informatique, j'ai personne à qui en parler. J'avais un pote qui était geek avec moi jusqu'à la fin du lycée, mais après il a fait une école de commerce et il a arrêté avec ces trucs-là. En revanche, pour ce qui est de l'intérêt intellectuel dans ce domaine, il y a pleins de gens avec qui tu peux parler de ces choses-là. Mon pote Quentin, par exemple : hier, on s'est fait un Mc Do, je travaillais en même temps, on regardait PSG-Barça. Le sandwich, l'ordi et le match. Quentin, il a vu un reportage « On n'est pas couchés » dans lequel ils ont parlé du fait que Google nous traque, etc. Je suis d'accord qu'il y a du souci à se faire. Le problème, c'est que beaucoup de gens parlent sans avoir un avis vraiment nuancé et référencé. Et mon pote Quentin, il disait que c'est scandaleux, bref, le petit outrage habituel que les gens ont quand ils entendent ça. Et je lui ai dit d'imaginer le coût de fonctionnement d'une boîte comme Google à la journée : ça se compte en dizaines de millions de dollars. Les mecs, ils ont des stades de foot entiers remplis de disques durs, quoi. Et tout est gratuit. Google, c'est tellement utile dans ta vie, et on te demande pas un centime. Et après tu t'offusques que les mecs te fassent des pubs. Mais

c'est leur *business model*, c'est leur seule source de revenus. Donc les pubs sur Internet, c'est des impôts ! Ça fait chier, mais c'est comme ça que Google peut te fournir le service qu'à la base t'es allé chercher. Donc moi je pense qu'il faut pas trop s'en plaindre quand même. Si c'est que ça, c'est pas méchant.

2. Enjeux pour l'analyse : travail en ligne et sociologie du travail

Il serait vain dans le cadre d'une postface d'énumérer les innombrables questions et hypothèses que peut soulever un tel entretien. D'ailleurs, que faire de ce témoignage, restitué ici sans connexion à un quelconque corpus de données empiriques ? Si l'on peut « penser par cas » (Passeron et Revel, 2005), il serait sans doute abusif d'inférer des lois générales à partir de la situation singulière considérée ici. Encore peut-on s'appuyer sur le discours d'Émilien et sur les informations factuelles qu'il rapporte pour voir dans quelle mesure le travail en ligne, en tant que nouveau terrain d'enquête, apporte du neuf à la sociologie du travail, ou, à l'inverse, fait valoir l'intérêt des acquis fondamentaux de ce domaine de recherche.

2.1. L'enjeu du collectif

Émilien définit son travail par la négative : il fait, dit-il, ce que ne font pas les algorithmes[9]. Par la formule *human intelligence tasks* convoquée dès le début de la conversation, il le dit en creux : c'est en référence à l'expression *artificial intelligence* que le travail des *raters* est qualifié d'humain. Parler d'« intelligence humaine » semble contradictoire avec ce qu'il dit de ces tâches, certaines d'entre elles pouvant être réalisées en étant « totalement absent intellectuellement ». Ce qui explique peut-être son indifférence ou la distanciation qu'il opère à l'égard du contenu de l'activité. Son intérêt n'est en effet ni dans le travail, ni dans les relations sociales au travail (elles sont quasi inexistantes), mais dans les conditions matérielles de sa réalisation. La flexibilité, présentée comme principale satisfaction avec la rémunération, est alors valorisée sous l'angle de la liberté de mouvement et de l'autonomie temporelle[10]. Mais cette autonomie spatio-temporelle du travail en ligne va de pair avec une autre de ses propriétés, le délitement du collectif jusqu'à son absence. Doublement à distance, Émilien est à la fois distancié du capital (son activité est déconnectée de l'espace physique de l'entreprise) et distancié du travail (il n'existe aucun lieu physique permettant l'émergence d'un collectif[11]). Comment avoir prise sur la régulation du travail dans une situation d'atomicité et de dispersion extrêmes des travailleurs ? Les « collègues » d'Émilien, qu'il ne connaît pas et dont il ignore le nombre, sont éparpillés dans le monde entier,

9. Il est d'ailleurs intéressant qu'il insiste davantage sur le « travail de complément » qu'il fait *après* l'algorithme que sur le travail qu'il fait *sur* l'algorithme. Or sa fonction officielle est d'évaluer la pertinence des résultats produits par les algorithmes de Google.
10. On observe des schèmes analogues de valorisation de l'autonomie temporelle et du nomadisme dans des univers caractérisés par un « travail sans qualités » (Sennett, 2000), par exemple chez les vacataires de l'industrie des sondages (Caveng, 2011) ou chez certains intérimaires (Glaymann, 2010).
11. Les modalités les plus probables d'interaction entre *raters* sont celles du forum de discussion sur Internet.

conformément au fantasme économiciste d'un marché du travail autorégulé, mondial et instantané. Tous les facteurs semblent donc réunis pour rendre improbable l'émergence d'un collectif de travailleurs. Pourtant, des épisodes récents de mobilisation collective montrent que la coopération entre travailleurs[12] et les conflits du travail[13] trouvent des conditions pour s'exprimer. La conjoncture est également propice, puisque la médiatisation croissante des enjeux de l'« ubérisation » du travail remet en discussion la question du « salariat déguisé » à laquelle Émilien fait explicitement référence, et qui pourrait fournir une base pour une mobilisation collective des *raters*. Les médias rapportent d'ailleurs que des travailleurs des géants de l'économie collaborative (Uber, Airbnb) peuvent parvenir, comme ce fut le cas en Californie, à faire requalifier leur prestation de service en contrat de travail salarié. Il existe donc des conditions pour qu'émergent, chez les travailleurs en ligne, des formes d'action collective. Mais quelles modalités peuvent-elles suivre, étant donné la forte dispersion géographique des travailleurs ? Il existe un ancrage territorial du droit du travail, mais les seules modalités d'interaction entre *raters* des quatre coins du monde sont télématiques. L'action collective des travailleurs en ligne est-elle alors condamnée à s'exprimer uniquement sur des supports, eux aussi, en ligne ? Quelle est la portée émancipatrice du Web et des réseaux sociaux pour ces derniers ? La rareté des cas observables ne permet pas encore d'apprécier la pertinence des concepts et des théories qu'a pu fournir jusqu'ici la sociologie des mouvements sociaux et des relations professionnelles.

2.2. L'organisation sociale du travail

Émilien dit de son travail qu'il correspond sans ambiguïté à du *crowdsourcing*, c'est-à-dire à du travail massivement distribué[14]. Dans la même veine, certains *raters* se définissent comme des *crowd workers*, des « travailleurs de la foule ». Ces termes nouveaux font oublier que la formule associant télématique et délocalisation à domicile existe depuis les années 1960[15]. Mais le système d'atomisme productif de SuccessTech semble introduire de nouveaux bouleversements face aux normes du travail salarié : micro-segmentation des tâches, distribution de ces tâches au clic[16], fractionnement du temps de travail, mesure des durées d'exécution à la seconde près et rémunération du travail par rubriques de tâches. L'émiet-

12. Les travaux de Kingsley *et al.* (2014) montrent que les *crowd workers* de la plate-forme de micro-travail d'Amazon collaborent *via* l'échange de tuyaux et de conseils sur des forums tels que www.mturkcrowd.com ou www.turkernation.com.

13. Voir par exemple les luttes hébergées sur le site www.wearedynamo.org.

14. Il illustre en cela l'adoption à plus grande échelle d'une acception nouvelle du terme *crowdsourcing*, qui initialement renvoyait à des processus d'externalisation de tâches complexes à des bénévoles. Cf. Karën Fort, « Miracles et mirages du *crowdsourcing* », *Libération*, 7 mai 2015 et *Le Journal du CNRS*, 7 mai 2015 :https://lejournal.cnrs.fr/billets/miracles-et-mirages-du-crowdsourcing.

15. Michel Lallement (1990, p. 200) fait mention d'une entreprise britannique appelée F. International Ltd, qui mobilise à partir de 1962 des centaines de programmateurs à domicile.

16. « C'est pas *mon* panier, c'est le panier de tout le monde en fait. » Ce système de distribution des tâches au clic est davantage une sorte d'avatar dématérialisé de la place de grève qu'un équivalent du « panier » du consommateur en ligne.

tement du travail poussé à son paroxysme va de pair avec une nouvelle métrologie du travail, dont le sens reste à élucider. Le dispositif de captation de travail de SuccessTech est basé sur une rémunération en temps de travail, contrairement au *Mechanical Turk*[17] d'Amazon, emblème contemporain du travail à la pièce. Mais, à regarder de près le système de SuccessTech, on s'aperçoit qu'à chaque tâche est attribué un *average expected time*, qui correspond à une somme d'argent bien précise. La mesure du temps n'est qu'une opération intermédiaire d'un calcul dont l'élément de base reste la tâche. De même, ce sont les tâches, et non un contrat de louage de main d'œuvre, qui font l'objet des échanges sur le marché du travail de cette sous-traitance à grande échelle. La métrologie de SuccessTech, à travers l'imposition d'une cadence, masque donc autant qu'elle consacre un retour en force du travail à la tâche. Le travail en ligne pour SuccessTech fait alors coexister deux modalités d'organisation du travail : un système planétaire de mobilisation de main d'œuvre et des pratiques de segmentation infra-individuelle des actes de travail. Le recul que procure l'histoire des formes de rémunération (Mottez, 1967) est une ressource précieuse pour replacer sur le temps long cette situation mêlant à la fois le contrat de louage du XIX^e^ siècle et le micro-travail mondialisé du XXI^e^ siècle. Tout aussi précieuses sont les études consacrées à l'histoire du travail à domicile (Lallement, 1990), en ce qu'elles invitent à mettre en rapport la situation des *raters* avec celle des ouvriers tisseurs de la soierie lyonnaise, des ouvrières de la dentelle ou de la fleur artificielle.

2.3. La mise en données du travail

Comme toute activité connectée (Beuscart *et al.*, 2009), le travail en ligne produit des données en quantité. Le travail d'Émilien est mesuré sous divers angles *via* des techniques de *time-tracking* : durée de chaque tâche, vitesse d'exécution, plages horaires des sessions de travail dans la journée, distribution du travail par type de tâche. Il existe donc des possibilités pour l'employeur de s'appuyer sur les traces numériques des travailleurs pour les contrôler par le biais d'un encadrement automatisé. Mais ce *quantified work* peut également induire, comme c'est le cas chez les préparateurs de commande d'Amazon ou des *drives* de supermarchés, un régime de mise au travail fondé sur la ludification du travail (Mollick et Rothbard, 2014). Une batterie d'indicateurs mesurant les performances en continu peut avoir des effets performatifs sur le travailleur, lorsque, voyant sa performance du jour comme la norme minimale à atteindre le lendemain, celui-ci est conduit à redéfinir constamment ses objectifs de performance[18]. Dans le même temps, le travail en ligne rend invisible tout un ensemble d'activités et de comportements chez les travailleurs qui, dans une situation en présentiel, pourraient

17. Sorte de marché électronique du travail, le *Mechanical Turk* est une application en ligne dans laquelle des tâches, pour la plupart peu complexes, sont proposées par des employeurs (entreprises, particuliers) à toute personne de plus de 18 ans disposant d'un ordinateur. Amazon effectue un prélèvement de 10 % sur l'ensemble des transactions. Voir Pierre Lazuly, « Télétravail à prix bradés sur Internet », *Le Monde diplomatique*, août 2006, n° 629, p. 16-17.
18. Cette sorte de fuite en avant est identifiée par la Commission nationale de l'informatique et des libertés / CNIL (2014) comme un risque inhérent aux pratiques de quantification *via* les capteurs numériques, qui « inscri[vent] les individus dans des processus de perfectionnement dont l'objectif recule au fur et à mesure qu'ils progressent ».

faire l'objet de contrôles par des supérieurs ou par des pairs. Que devient alors l'évaluation du travail dans ce type de configuration ? Se réduit-elle à une simple évaluation des *outputs* ? S'éloigne-t-elle radicalement des formes plus « classiques » d'évaluation du travail au point de se rapprocher davantage, par exemple, de celle qui prévaut dans les MOOCs[19] ? Dans quelle mesure les entreprises de sous-traitance en masse s'appuient-elles sur ces métriques pour classer les travailleurs, pour organiser l'éviction des moins productifs, pour récompenser les plus performants ? Traiter ces questions supposerait de franchir le miroir sans tain qui sépare Émilien de SuccessTech, et observer les pratiques, les logiques et les paradoxes propres à ces plates-formes de micro-travail.

Émilien et ses homologues semblent travailler dans un laboratoire du post-salariat. Mais les *raters* doivent surtout être resitués dans la nébuleuse plus vaste des travailleurs en ligne. Le monde des *crowd workers* compte des milliers de travailleurs payés à la tâche pour *tagger* des images et des vidéos, classer des contenus, faire de la reconnaissance de texte, traduire ou réécrire des fragments de texte, faire de la retranscription d'enregistrements oraux, etc. (Smith et Leberstein, 2015). Mais l'industrie de la donnée engendre, on l'a vu tout au long de cet ouvrage, d'autres formes de travail extra-salarial. Il faudrait, pour compléter le tableau, resituer le travail en ligne dans le champ du *digital labor* (Cardon et Casilli, 2015), qui comprend notamment le fameux « travail invisible » des utilisateurs du Web et des objets connectés. Ces deux formes de contribution économique aux entreprises du numérique ont-elles suffisamment en commun pour être subsumées dans une même abstraction ? Le « travail de fourmi » d'Émilien n'est-il pas davantage un « travail » que nos activités quotidiennes sur Internet ? Cette question toute naïve a au moins le mérite de souligner l'arbitraire de la catégorie « travail », dont les contours sont renégociés au rythme des évolutions du système productif et des rapports sociaux (*Tracés*, 2015). Au-delà, elle invite à décomposer la nébuleuse du *digital labor* pour distinguer, comme le propose Marie-Anne Dujarier (2016), les différentes modalités de mise au travail imaginées par les plates-formes numériques. Mais, en définitive, l'enjeu n'est pas tant de savoir s'il faut développer une sociologie du travail en ligne, une sociologie du *digital labor*, ou une sociologie des plates-formes numériques. Il est de savoir comment, à rebours d'une tendance au morcellement des objets d'étude, prendre appui sur les questions fondamentales que se posent les sciences sociales au sujet du travail. Ce qui nous ramène, là aussi, au XIX^e^ siècle.

19. Les *massive open online courses* sont des programmes de formation en ligne auxquels peuvent s'inscrire plusieurs milliers de participants, qui sont ensuite évalués par le moyen de tests standardisés.

Références

Bastard I., Cardon D., Fouetillou G., Prieur C. et Raux S. (2013), « Travail et travailleurs de la donnée », *InternetActu.net*, http://www.internetactu.net/2013/12/13/travail-et-travailleurs-de-la-donnee.

Beuscart J.-S., Dagiral E. et Parasie S. (2009), « Sociologie des activités en ligne », *Terrains & travaux*, vol. 1, nº 15, p. 3-28 (https://www.cairn.info/revue-terrains-et-travaux-2009-1-page-3.htm).

Cardon D. (2013), « Dans l'esprit du PageRank. Une enquête sur l'algorithme de Google », *Réseaux*, nº 177, p. 63-95 (https://www.cairn.info/revue-reseaux-2013-1-page-63.htm).

Cardon D. et Casilli A. (2015), *Qu'est-ce que le Digital Labor ?*, Paris, Ina Editions, coll. « Études et controverses ».

Caveng R. (2011), *Un laboratoire du « salariat libéral » : les instituts de sondage*, Paris, Éditions du Croquant.

Commission nationale de l'informatique et des libertés (2014), « Le corps, nouvel objet connecté du *quantified self* à la m-santé : les nouveaux territoires de la mise en données du monde », *Cahiers innovation & prospective*, nº 2.

Dujarier M.-A. (2016), « Digital labor, travail du consommateur : quels usages sociaux du numérique ? », *Ina global*, http://www.inaglobal.fr/numerique/article/digital-labor-travail-du-consommateur-quels-usages-sociaux-du-numerique-8729?tq=7.

Glaymann D. (2010), *L'Intérim*, Paris, La Découverte, coll. « Repères ».

Kingsley S.C., Gray A.M.L. et Suri S. (2014), « Monopsony and the crowd: labor for lemons? », *IPP2014: Crowdsourcing for Politics and Policy*, Oxford Internet Institute, University of Oxford.

Lallement M. (1990), *Des PME en chambre : travail et travailleurs à domicile d'hier et d'aujourd'hui*, Paris, L'Harmattan.

Mollick E.R. et Rothbard N. (2014), « Mandatory fun: consent, gamification and the impact of games at work », *The Wharton School Research Paper Series*: http://ssrn.com/abstract=2277103.

Mottez B. (1967), « Formes de salaire et types d'action ouvrière », *Le Mouvement social*, nº 61, p. 5-12.

Passeron J.-C. et Revel J. (2005), *Penser par cas*, Paris, Éditions de l'EHESS.

Sennett R. (2000), *Le Travail sans qualités : les conséquences humaines de la flexibilité*, Paris, Albin Michel.

Smith R. et Leberstein S. (2015), *Rights on Demand: Ensuring Workplace Standards and Worker Security in the On-Demand Economy*, New York, National Employment Law Project.

Tracés (2015), « Appel à contributions au nº 32 : *Déplacer les frontières du travail* » (https://traces.revues.org/6378).

Diffusion-distribution :
FMSH-diffusion/CID
18 rue Robert-Schuman
94227 Charenton-le-Pont Cedex
Fax :(0033-1) 53 48 20 95

Impression d'après documents fournis
bialec, heillourt (France)
Dépôt légal n° 92022 - octobre 2017

Conférences du Collège de France

Big data et traçabilité numérique

Les sciences sociales face à la quantification massive des individus

sous la direction de

Pierre-Michel Menger et Simon Paye

Les traces numériques de l'activité des individus, des entreprises, des administrations, des réseaux sociaux sont devenues un gisement considérable. Comment ces données sont-elles prélevées, stockées, valorisées, et vendues ? Et que penser des algorithmes qui convertissent en outil de contrôle et de persuasion l'information sur les comportements, les actes de travail et les échanges ? Les *big data* sont-elles à notre service ou font-elles de nous les rouages consentants du capitalisme informationnel et relationnel ? Les sciences sociales enquêtent sur les enjeux sociaux, éthiques, politiques et économiques de ces transformations. Mais elles sont elles aussi de plus en plus consommatrices de données numériques de masse. Cet ouvrage collectif explore l'expansion de la traçabilité numérique dans ces deux dimensions, marchande et scientifique.

L'ouvrage est dirigé par Pierre-Michel Menger, professeur au Collège de France et titulaire de la chaire de « Sociologie du travail créateur », et par Simon Paye, maître de conférences à l'université de Lorraine, sociologue du travail et des groupes professionnels.

L'édition électronique de cet ouvrage est disponible sur :
https://books.openedition.org/cdf/4987

22 €
ISBN : 978-2-7226-0466-7

COLLÈGE DE FRANCE
1530