Des données du web pour faire de la sociologie... du web ?

Jean-Samuel Beuscart, Simon Paye, Pierre-Michel Menger

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

Pierre-Michel Menger et Simon Paye (dir.)

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

Édition imprimée
Date de publication : 24 octobre 2017
ISBN : 9782722604667
Nombre de pages : 218

http://books.openedition.org

Référence électronique
MENGER, Pierre-Michel (dir.) ; PAYE, Simon (dir.). *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus*. Nouvelle édition [en ligne]. Paris : Collège de France, 2017 (généré le 24 octobre 2017). Disponible sur Internet : <http://books.openedition.org/cdf/4987>. ISBN : 9782722604674.

Big data et traçabilité numérique

Les sciences sociales face à la quantification massive des individus

Jean-Samuel Beuscart <jsbeuscart@gmail.com>

Jean-Samuel Beauscart <jsbeauscart@gmail.com>

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

Jean-Samuel Beauscart <jsbeauscart@gmail.com>

Jean-Samuel Beauscart <jsbeauscart@gmail.com>

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.

Jean-Samuel Beauscart <jsbeauscart@gmail.com>