LES ENJEUX de l'information et de la communication

Dossier 2018

REVUE SCIENTIFIQUE EN SCIENCES DE L'INFORMATION ET DE LA COMMUNICATION

Numéro 2/2018 - Dossier thématique

Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication

Ce dossier a été coordonné par Vincent Bullich et Viviane Clavier.

Avec la participation de : Caroline Creton - Gabrielle Silva Mota Drumond - Alexandre Coutant - Florence Millerand - Anaïs Theviot - Philippe Chevallier - Jean-Marc Francony - Valentyna Dymytrova - Jean-Sébastien Vayre - Olivier Koch - Alexandre Joux - Marc Bassoni - Anne Lehmans - Guillaume Sire - Julien Rossi - Jean-Edouard Bigot

GRESEC
UNIVERSITÉ Grenoble Alpes

https://lesenjeux.univ-grenoble-alpes.fr

Table des matières

Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication

Présentation du dossier

Production of data, "Production of Society". Big Data and Algorithms in the field of Information and Communication Sciences

Producción de datos, "Producción de la sociedad". Big Data y Algoritmos en el campo de las Ciencias de la Información y de la Comunicación

Article inédit, mis en ligne le 15 novembre 2018.

Vincent Bullich

Vincent Bullich est maître de conférences en sciences de l'information et de la communication et membre du Groupe de Recherche sur les Enjeux de la Communication (GRESEC – Université Grenoble Alpes). Ses travaux concernent principalement l'analyse socio-économique des industries culturelles et communicationnelles ainsi que l'économie politique de la propriété intellectuelle.

Viviane Clavier

Viviane Clavier est maître de conférences HDR au Groupe de Recherche sur les Enjeux de la Communication (Gresec) à l'Université Grenoble Alpes. Ses recherches s'inscrivent dans le champ de l'information spécialisée et professionnelle et portent plus particulièrement sur l'organisation des connaissances dans les dispositifs info-communicationnels.

Le dossier 2018 des *Enjeux de l'information et de la communication* est consacré au couple « *Big Data* et algorithmes », termes récemment devenus prégnants dans les discours sur « la révolution numérique ». Portés tant par des journalistes, des essayistes, des acteurs politiques ou économiques que des représentants de la société civile, ces discours, tantôt apologétiques, tantôt apocalyptiques, opposent souvent promesses d'un renouveau économique aux risques majeurs pour la démocratie et la sécurité nationale, les libertés individuelles, voire l'avenir de l'espèce humaine que feraient

courir la captation et l'exploitation algorithmiques de volumes massifs de données. Concomitamment à cette mobilisation « triviale », le sujet s'est, sur le plan scientifique, échappé des cénacles des spécialistes en informatique pour s'imposer comme thématique des sciences humaines et sociales.

Au-delà des considérations méthodologiques relatives à une nouvelle heuristique fondée sur l'obtention et le traitement d'une quantité inédite de données (Anderson, 2008 ; Aiden et Michel, 2014) et que nous laisserons ici de côté, ce qui motive l'intérêt des chercheurs en SHS se rapporte aux mutations sociales imputables à ces nouveaux dispositifs « *Big Data* et traitement algorithmique ». En effet, ceux-ci seraient en passe d'investir de multiples sphères professionnelles et domestiques et configureraient un nombre croissant d'activités tant publiques que privées. Illustrant le phénomène, le néologisme « data-ification » s'impose progressivement dans les entreprises et les administrations afin de désigner « le nouvel âge » des systèmes d'information au sein de ces organisations (Delort, 2015). Cet essor de la production et du traitement des données est particulièrement visible au niveau des services de marketing ou des ressources humaines notamment (*Ibid.*), mais également dans le cadre des « stratégies de production de l'information publique » (Bardou-Boisnier et Pailliart, 2012) et du « gouvernement par les nombres », selon l'expression d'A. Desrosières (2008), conduits par l'Etat et ses administrations. En cela, le couple « *Big Data* et algorithmes » apparaît symptomatique de l'accentuation rapide et significative de « l'informationnalisation » de la société, c'est-à-dire de la circulation croissante et accélérée des flux d'information, ainsi que leur contribution accrue à l'ensemble des dimensions de la vie sociale (Miège, 2004). La dynamique est déjà ancienne et fut initialement abordée par le prisme exclusif du rapport de l'information à la production : elle apparaît, par exemple, à la fois comme la cause et la conséquence de l'émergence de la « société de l'information » comme projet industriel et projet de société à partir des années 1970-1980 (Mattelart, 2001). Cependant, depuis 2001, on assiste à une explosion sans précédent du volume de données médiatisées, explosion qui serait donc à même de provoquer des mutations majeures, en cours ou à venir, à laquelle l'ensemble des acteurs sociaux doivent se préparer, afin d'en tirer le meilleur parti ou d'organiser la résistance à ce qui pourrait être un « panoptique » au niveau mondial (Kitchin, 2014).

Malgré l'ampleur récente des discours profanes et experts oscillant entre promesses exubérantes et anathèmes radicaux et malgré l'intérêt nouveau des chercheurs en SHS pour la thématique des *Big data* et algorithmes, la production même de ces données massives et des instruments de traitement algorithmique reste encore peu appréhendée comme objet de recherche par les sciences humaines et sociales. C'est donc sur cet aspect que le dossier 2018 des *Enjeux de l'information et de la communication* a souhaité se centrer. Plus précisément, celui-ci a visé à une meilleure connaissance du procès de production des *Big Data* et algorithmes en tant que fondement de dispositifs décisionnels de « production de la société ».

La référence à l'ouvrage d'A. Touraine (1973) n'est pas fortuite : bien que la démarche ne se veuille pas (nécessairement) sociologique, il s'agit bel et bien d'envisager ces dispositifs comme des « instruments de production de la société par elle-même », « instruments » foncièrement « historicisés », c'est-à-dire à la fois reflets d'une époque et moteurs de l'évolution sociétale à cette période (*Ibid.*). Nous faisons donc, en filigrane, l'hypothèse que la compréhension des effets sociaux (*lato sensu*) de ces dispositifs passe immanquablement par la compréhension de la manière dont ces données sont construites, traitées et utilisées. Dans cette perspective, l'analyse du procès concret de production des dispositifs constitue une étape indispensable à la connaissance du type de société qu'ils seraient à même de produire. Or, cette analyse passe par l'étude des instruments et techniques mobilisés, par celle des activités des intervenants qui les conçoivent et les mettent en œuvre, par la prise en compte des discours qui orientent et accompagnent leur ancrage social, ainsi que par l'éclairage des conditions institutionnelles dans lesquelles il se réalise. Il s'agit ainsi de

refléter l'épaisseur des multiples médiations *qui s'exercent sur* ces dispositifs, et *qu'exercent ces* dispositifs dans la configuration du rapport des individus aux mondes sociaux. C'est précisément à cet ensemble de domaines que se rattachent les travaux qui composent ce dossier. Ceux-ci ont été répartis en quatre axes, en fonction des objets de recherche et positionnements scientifiques de leur(s) auteur(s), axes dont la complémentarité est à même de favoriser la compréhension des enjeux liés à la production sociale telle que configurée par la production de données.

Les *Big Data* au prisme des stratégies industrielles et marchandes

Ce premier axe se rapporte à l'étude des stratégies des acteurs économiques, aux logiques de production industrielle des données et de leur traitement, dès lors envisagés comme ressources, ainsi qu'à l'étude de la construction des marchés sur lesquels ces mêmes ressources sont valorisées. Au cœur de la démarche, se trouve l'identification des positionnements de ces acteurs et l'analyse des rapports qu'ils entretiennent. Il s'agit, en outre, d'aborder les données massives et algorithmes comme des facteurs de production (à l'origine de la désormais fameuse « *Data-Driven Economy* ») au sein de filières dont le fonctionnement reste à éclairer et des ressources valorisables *per se* sur des marchés. Enfin, cet axe s'intéresse aux discours portés par ces acteurs et contribuent fortement à inscrire ce couple *Big Data* et algorithmes dans des « économies de la promesse » au sein desquelles il est présenté comme « ressource essentielle » et « solution ultime » aux défis économiques et sociaux de demain (Bullich, 2016).

Considérant l'opacité qui règne dans « l'économie numérique » et la culture du secret industriel qu'y manifestent les acteurs industriels (*a fortiori* les acteurs les plus importants), il faut généralement prendre des chemins détournés afin d'observer la conduite de ces mêmes acteurs industriels. C'est précisément la démarche qu'emprunte Caroline Creton. Elle s'intéresse à la façon dont Facebook est à même de valoriser auprès des musiciens les données dont l'entreprise dispose sur les goûts culturels des utilisateurs de son dispositif. La stratégie de l'entreprise californienne est ainsi mise en lumière par un travail sur les pratiques et usages de musiciens « à notoriété locale », leurs « tactiques », suivant la terminologie de de Certeau, pour répondre aux logiques commerciales de la plateforme et la compréhension qu'en ont ces usagers. Ceux-ci se trouvent face au dilemme que l'auteur exprime dans le titre de l'article : faut-il payer ou non pour toucher son public ? La pratique rappelle le « *pay per play* » pratiqué aux Etats-Unis dès le début du vingtième siècle par certaines salles de concerts et producteurs de spectacles et qui consiste à faire payer aux musiciens le droit de jouer sur scène en arguant de la force promotionnelle du lieu. Il en va de même en 2018 mais ce sont les données personnelles dont dispose Facebook qui font ici figure d'argument commercial décisif. Caroline Creton montre qu'au-delà de l'enjeu économique, il s'agit pour ces musiciens d'acquérir des compétences de l'ordre du marketing numérique afin d'utiliser au mieux les outils de la plateforme. En cela, « l'offre industrielle de Facebook participe aux mutations qui font du musicien un artiste-entrepreneur » (Creton, 2018).

Le deuxième article étudie les stratégies liées aux données massives d'un autre acteur majeur de la Silicon Valley : Netflix. Gabrielle Silva Mota Drumond, Alexandre Coutant et Florence Millerand proposent ainsi d'étudier les ressorts du dispositif de recommandation qui a fait le succès – tant médiatique qu'économique – de l'entreprise. La thèse est que ce dispositif serait à même de « produire l'usager » ainsi que l'indique le titre de l'article. Afin d'étayer leur démonstration, les auteurs ont mis en place une méthodologie fondée principalement sur l'étude de la mise en visibilité des contenus de Netflix ainsi que sur les discours de l'entreprise sur son propre système de recommandation. Ce qui apparaît en effet remarquable est que celle-ci, au travers de son « *Techblog* », communique de façon régulière à ses différents publics sur les choix opérés par l'entreprise quant à ce dispositif censé « optimiser » l'expérience utilisateur. Or, cette optimisation

est ramenée, ainsi que nous l'indique les auteurs, à une consommation intensive, « excessive », de contenus audiovisuels (Drumond, Coutant, Millerand, 2018).

Le troisième et dernier article se rapporte quant à lui au champ politique et s'intéresse à la prégnance toute contemporaine des *big data* dans les stratégies électorales. Anaïs Théviot présente ainsi les stratégies des « prestataires en *big data* électoral » et retrace l'émergence puis l'affirmation d'une « croyance en l'efficacité » des données massives, croyance désormais partagée avec les décideurs politiques et militants. Ce que l'auteur met en évidence est que le moteur de cette technicisation des campagnes électorales réside dans la « promesse » portée par ces entrepreneurs en données quant aux possibilités de prédiction, d'orientation des stratégies électorales qu'une connaissance des électeurs serait en mesure de produire (Théviot, 2018). Ces discours de promotion ont eu une portée perlocutoire considérable puisque le recours aux données massives est maintenant perçu par les responsables des équipes électorales comme indispensable dans la panoplie des outils de campagne. En cela, l'article d'Anaïs Théviot illustre parfaitement les effets sociaux (en l'occurrence politiques même) produits par les stratégies des acteurs économiques autour de l'obtention et du traitement des *Big Data*.

La production des *data* : dispositifs et pratiques

Le deuxième axe concerne l'analyse du procès de production lui-même. Il s'agit ainsi d'ouvrir la « boîte noire », la focale se faisant plus resserrée. Les données sont souvent considérées comme « brutes », comme si elles étaient uniquement le point de départ de traitements algorithmiques, alors qu'elles sont « toujours déjà » le résultat de traitements élaborés, qu'elles portent sur des gisements informationnels ou documentaires volumineux préalablement construits (archives institutionnelles ou patrimoniales, bibliothèques numériques par exemple). Il s'agit dès lors de s'interroger à la fois sur la provenance des données mobilisées, sur leur transformation en informations, sur le « travail » des agents impliqués dans cette production ainsi que sur les modalités organisationnelles qui président à leur co-opération. C'est donc le *process* lui-même qui est ici au cœur de l'investigation : il s'agit d'appréhender à la fois empiriquement et théoriquement les différentes phases qui le composent, de l'obtention des data comme *inputs* à la production spécifique d'*outputs*, résultat, généralement automatisé, du traitement algorithmique. En outre, la transformation qui y est opérée doit être abordée sous l'angle de ce qui se perd, ou au contraire ce qui est ajouté, en termes informationnels, par le *process* même. Dans le cadre de ce deuxième axe, il s'agit donc de se centrer à la fois sur des dispositifs comme ensemble d'instruments et de techniques ainsi que sur des pratiques, fondées sur des savoirs et savoir-faire. Il s'agit d'interroger aussi les formes de (re-)contextualisation des *data*, les modalités de leur interprétation, la place des langages et celle des représentations visuelles.

En 2006, Bruno Maresca soulignait que le recours systématique aux enquêtes de fréquentation des bibliothèques de lecture publique en France faisait « apparaître ce domaine de la culture comme l'un des plus préoccupés de la quantification des publics ». En s'intéressant aux méthodes de recueil de données destinées à documenter les usages de la bibliothèque numérique Gallica, Philippe Chevallier (2018) s'inscrit à la fois dans une continuité, celle des études de publics, et une rupture, celle de l'incursion de méthodes complexes déléguées à des spécialistes de la fouille de données. Dans cette première contribution, Philippe Chevallier évoque comment les données de masse recueillies sous forme de logs, peuvent renseigner finement sur la temporalité des connexions (fréquence, durée) des usagers, sur leurs actions-types, sur leurs parcours de navigation ou encore sur les types de documents consultés. Il indique également que ces méthodes sont complémentaires d'autres enquêtes et observations qualitatives en vigueur à la BnF et met en garde

sur la nécessité d'une approche concertée de la donnée entre professionnels de bibliothèques et *data scientists* à des fins de gouvernance.

Le deuxième article proposé par Jean-Marc Francony (2018) s'intéresse aux flux de données publicisés dans les API de Twitter et porte un regard expert sur la fabrication des données et les logiques de publication. Tirant profit de sa notoriété, Twitter, qui figure parmi les dispositifs les plus utilisés, a récemment repensé son modèle économique en enrichissant l'accès à ses interfaces de données. Parallèlement, de nombreuses restrictions qualitatives et quantitatives ont été opérées sur les données, cette réduction informationnelle fondée sur des contraintes techniques conduisant in fine à réduire les possibilités d'accès aux réseaux d'acteurs et aux flux informationnels. Ce faisant, Twitter préserve et garde la maîtrise de son offre de service.

Enfin, l'axe se termine par la contribution de Valentyna Dymytrova (2018) qui focalise son attention sur les médiations accompagnant l'ouverture des données numériques urbaines. A partir d'une sélection d'applications mobiles destinées à assister les usagers dans leurs déplacements en ville, l'auteur distingue plusieurs types de médiations, notamment informatiques et info-communicationnelles. Valentyna Dymytrova montre de manière très convaincante qu'il est indispensable de considérer les conditions de production des applications, la nature des acteurs, les moyens matériels, technologiques et symboliques mobilisés pour révéler et éclairer les médiations techniques à l'origine des choix des développeurs. Ainsi, cette étude qui s'appuie sur les résultats d'une enquête menée en France dans le cadre d'une ANR (OpenSensingCity), illustre parfaitement notre propos. Elle montre que les nouveaux services urbains qui reposent sur la circulation et la valorisation des données sont toujours le résultat « d'une double médiation technique et sociale et reflètent des orientations des acteurs dominants qui en maîtrisent la production et la distribution ». (*Ibid.)*

L'intégration des *Big Data* et algorithmes au sein des secteurs professionnels et des champs sociaux

Le troisième axe rassemble des analyses sectorielles. Il s'agit de s'intéresser ici aux modalités d'intégration des dispositifs au sein de secteurs professionnels et de champs sociaux : assiste-t-on à une reconfiguration des activités qui se déploient en leur sein ou, plus modestement, à des mutations partielles de celles-ci ? Comment de grands ensembles d'activités comme la culture, le marketing, la santé, le journalisme ou l'éducation sont-ils affectés de façon spécifique et/ou de façon commune, comparable ? Afin d'apporter des éléments de réponses à ces questions, il s'agit de porter ici une attention plus soutenue aux professionnels qui participent à cette production. Les récentes fonctions aux dénominations anglophones qui se multiplient au sein des organisations telles que « *Data Scientist* », « *Data Analyst* », « *Chief Data Officer* » ou « *Data Protection Officer* », manifestent l'intérêt croissant des entreprises et des administrations pour l'obtention, l'agrégation et l'analyse de données massives préalablement à la conception et la mise en œuvre de leurs stratégies. Ces fonctions ainsi que leur intégration organisationnelle ont ainsi fait l'objet d'études spécifiques.

Premièrement, Jean-Sébastien Vayre s'est intéressé dans son article aux « technologies d'apprentissage artificiel appliquées à la gestion de la relation client ». Il a décortiqué minutieusement les dispositifs mis en place par différentes entreprises et un des points forts de l'étude est de révéler l'importance du « travail de cadrage de leurs activités inférentielles » (Vayre, 2018). Ce travail est le fruit de « *data scientists* » employés par les entreprises considérées. Au moyen d'une étude portant sur leur activité complétée par une analyse de l'architecture technico-logique de ces « machines apprenantes », l'auteur affirme que ces dispositifs sont à même, au

travers de ce qu'il nomme la « socialisation algorithmique » (c'est-à-dire des processus automatisés d'acquisition de compétences à la fois cognitives et communicationnelles), d'élaborer leurs propres critères de pertinence révélant leur capacité à produire, plus que reproduire, de l'information.

Le deuxième champ d'application abordé dans le cadre de cet axe est le domaine militaire. Olivier Koch propose ainsi un article exposant les modalités d'inclusion de machines prédictives, fondée sur l'obtention de données massives et leur traitement exhaustif *via* des algorithmes informatiques, au sein de l'armée des Etats-Unis d'Amérique. Depuis le début de la décennie 2000, l'Etat-Major américain s'est doté d'outils destinés à repérer et prévenir les foyers insurrectionnels sur les principaux terrains d'intervention où son armée s'est déployée (Afghanistan et Irak). Toutefois, au-delà du caractère directement opérationnel en zone de conflits et situation de « guerre irrégulière », ce que montre l'auteur est que l'intérêt de ce type de dispositif réside dans les possibilités (ou les « promesses » tout du moins) « d'optimisation de la prise de décision politique » (Koch, 2018). En effet, malgré « l'inefficacité des programmes mis en œuvre au regard des finalités que leur assignent leurs artisans-concepteurs », ces dispositifs continuent d'être financés et exploités précisément en raison, une nouvelle fois, d'une croyance « au plus haut niveau de l'Etat fédéral » dans l'efficace de leur prédictions, non pas sur le plan de la polémologie mais sur celui de la gouvernementalité (*Ibid*.).

Le dernier article de ce troisième axe se rapporte aux « résistances épistémologiques, ruptures économiques et adaptations professionnelles » liées à l'intégration des *Big Data* au sein des médias d'informations. L'insertion des dispositifs liés à l'acquisition et au traitement des *Big Data* dans les champs professionnels ne se fait pas sans heurts et c'est précisément ce qu'illustre l'article d'Alexandre Joux et Marc Bassoni. Spécialistes du journalisme, les deux auteurs exposent les principales mutations du métier liées à la « data-ification » des rédactions. Ils procèdent pour ce faire à une « cartographie des études » récentes sur le sujet et mobilisent également des paroles d'acteurs qui témoignent de l'attitude pour le moins ambivalente de certains professionnels face au « journalisme robotisé » (Joux, Bassoni, 2018). Au final, ce que les auteurs critiquent foncièrement est la « vision ultra-techniciste » qui présente le traitement des données massives comme un nouveau mode d'investigation journalistique qui se distinguerait par son objectivité totale, et en cela, une « prophétie de la disparition du métier » pour les plus pessimistes (*Ibid*.). Les faits militent pour une perspective (évidemment) plus nuancée, bien que les mutations à venir s'annoncent profondes.

Les processus de régulation : gouvernance, lois et réglementations, normalisation technique et organisationnelle

Le quatrième et dernier axe porte sur les questions de régulation du procès de production « *Big Data* / algorithme(s) de traitement ». Par régulation, nous entendons aussi bien les règles s'imposant « verticalement » aux acteurs sociaux que celles produites par ceux-ci, généralement dans une visée d'optimisation de la coordination des activités. Ainsi, le regard se pose-t-il tant sur les lois et règlements produits par l'Etat et ses administrations que sur les normes *ad hoc* configurant les dispositifs et les pratiques professionnelles. Il s'agit notamment de rendre compte de la plasticité des règles produites ainsi que des asymétries en termes d'agentivité entre producteurs de *data* et producteurs d'informations à partir de ces *data*, entre les acteurs maîtrisant cette production d'information et leurs concurrents ne la maîtrisant pas, asymétries que ces règles sont susceptibles de renforcer ou de réduire. Il s'agit, de surcroit, de porter le regard sur les processus de production des normes et les jeux argumentatifs, sur les instances de délibérations ainsi que sur les discours qui y sont énoncés, considérant que ces discours sont porteurs de visions

du monde à partir desquelles ces normes se construisent et qui les justifient, les légitiment aux yeux des acteurs sociaux.

Enfin, ce dernier axe concerne également les travaux s'intéressant à l'inscription de la thématique des *Big Data* et algorithmes dans les politiques publiques. Au-delà des questions portant sur l'encadrement légal *stricto sensu*, il s'agit d'apprécier comment l'Etat investit ce domaine, dans le cadre des mesures concernant l'*Open Data* évidemment, mais aussi au travers de ses actions favorisant le développement des activités liées à la captation et au traitement des données massives : dans le domaine de la formation, dans celui de la standardisation technique, etc.

L'article d'Anne Lehmans s'inscrit précisément dans cette optique. L'auteur s'intéresse aux « politiques et pratiques de médiation, de valorisation et d'éducation autour des données ouvertes ». A la suite d'une enquête collective « sur plusieurs entités productrices et utilisatrices de données ouvertes dans l'académie de Bordeaux » et dont l'objectif est une meilleure connaissance de « la façon dont les enseignants recourent effectivement à ces données ouvertes dans un objectif pédagogique », l'auteur montre que les mesures traduisant les politiques autour de l'*Open Data* sous-estiment généralement le poids des contraintes d'appropriation (Lehmans, 2018). Par conséquent, si la mise à disposition de données publiques est susceptible de contribuer à un élargissement de l'espace public, en favorisant notamment un dialogue renouvelé entre « les collectivités publiques, les entreprises et les citoyens », l'auteur insiste sur le fait que le rapport aux données nécessite des compétences, une « culture », dont ne dispose pas systématiquement chaque citoyen ni même chaque entreprise ou administration. Un effort de formation voire « d'acculturation » est donc nécessaire (*Ibid*.).

On retrouve les interrogations sur la normalisation dans l'article de Guillaume Sire. A partir d'une étude du Web sémantique, l'auteur expose comment « la discrétisation des informations présentes sur le Web, c'est-à-dire leur transformation en *Big Data* » contribue à industrialiser des « politiques du sens » (Sire, 2018). Pour ce faire, il compare trois syntaxes et procédures de normalisation afférentes et montre comment les pratiques des producteurs de contenus se sophistiquent dans la perspective d'un référencement optimisé. Pis, il identifie des « stratégies prédatrices » qui passent par l'imposition d'éléments syntaxiques que mènent les géants du Web, au premier desquels se trouve Google/Alphabet. Derrière des procédures de normalisation et de régulation pour plus d'interopérabilité, se produit une « tectonique […] dont dépendent les modalités concrètes de production et de mise en circulation à grande échelle des signes, c'est-à-dire des informations transformées en données qualifiables, puis qualifiées ». Or, ainsi que le pose l'article, cette « tectonique » manifeste avant toute chose des mouvements industriels, des affrontements entre « visions [...], projets, intérêts » divergents (*Ibid.*).

Le dernier article de ce quatrième axe renvoie à la question du droit à la protection des données personnelles au sein de la recherche. Dans leur contribution, Julien Rossi et Jean-Edouard Bigot indiquent que les chercheurs en SHS seraient actuellement mis au défi d'intégrer les équipements numériques pour mener leurs recherches. Ce faisant, les chercheurs seraient progressivement confrontés à divers types de données numériques (traces numériques, réponses à des questionnaires, entretiens, etc.) qui, en raison de leur caractère personnel, tomberaient sous le coup de la loi, notamment du Règlement général de protection des données entré en vigueur le 25 mai 2018. A la faveur d'entretiens semi-directifs conduits auprès de chercheurs en SHS engagés dans une ANR, les auteurs ont enquêté sur les formes d'autorégulation et sur les questionnements éthiques soulevés par ce cadrage juridique des données personnelles (Rossi, Bigot, 2018).

Conclusion

Les articles composant ce dossier 2018 des *Enjeux de l'information et de la communication* envisagent donc la question des données massives aux prismes des approches et des méthodologies multiples mobilisées en sciences de l'information et de la communication. Si la question de l'épistémologie n'est pas abordée de front, il ressort toutefois de ces 12 contributions un consensus fort quant à l'exigence scientifique de considérer ces données comme, avant toute chose, des construits sociaux. En cela, elles manifestent autant les perspectives, les positionnements, voire les stratégies de leurs (co-)concepteurs que les caractéristiques supposées « immanentes » des objets analysés ; ignorer leurs conditions de production revient donc à se priver d'un élément majeur de compréhension de leur fonctionnement et de leurs effets sociaux. Ces contributions éclairent également les modalités de production du social induites en retour par les traitements des données. Ce faisant, les différents articles exposent, chacun à leur façon, comment les dispositifs contribuent à produire ou reproduire des relations entre les individus ou alors les relations des individus aux institutions qui les gouvernent.

Au travers de ce parcours en quatre étapes, nous espérons présenter au lecteur un large panorama des modes de production « algorithmique » de la société, en alternant les points de vue, les focales et en multipliant les objets visés.

Avant de clore cette présentation, nous souhaitons remercier très chaleureusement l'ensemble des auteurs ayant contribué au dossier : pour la qualité de leurs travaux ainsi que pour leur grande réactivité lors des échanges que nous avons eus avec le comité éditorial.

Références bibliographiques

Aiden E. et Michel, J.-B., *Uncharted: Big Data as a Lens on Human Culture*, New York : Riverhead Books, 2014.

Anderson C., « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete», *Wired*, 2008 [en ligne] : http://www.wired.com/2008/06/pb-theory/ (consulté le 10/04/2017).

Bardou-Boisnier S. et Pailliart I. (coord.), dossier : *Information publique : stratégies de production, dispositifs de diffusion et usages sociaux*, *Les Enjeux de l'Information et de la Communication*, n° 13/2, 2012, [en ligne] : http://w3.u-grenoble3.fr/les_enjeux/pageshtml/art2012.html#dossier (consulté le 12/04/2017).

Bullich, V, « Big Data : stratégies industrielles et économie de la promesse. », dans Pilati A. (dir.), *La comunicazione multipla. Media, piattaforme digitali, Over the Top, Big Data*, Rome : Magna Carta Edizioni, 2016, p. 41-71.

Certeau, M. (de), *L'invention du quotidien. Tome 1. Arts de faire*, Paris : Gallimard, 2002 (1ère édition : 1980)

Chevallier, P., « Les données au service de la connaissance des usages en ligne :l'exemple de l'analyse des logs de Gallica », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Creton, C, « *To pay or not to pay* : les musiciens à notoriété locale face à la publicité ciblée sur Facebook », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Delort, P., *Le Big Data*, Paris : Presses Universitaires de France, 2015.

Desrosières A., *Gouverner par les nombres*, Paris : Presses des Mines ParisTech, 2008

Drumond, G., Coutant, A., Millerand, F, « La production de l'usager par les algorithmes de Netflix », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Dymytrovna V., « Les médiations de l'open data au prisme des applications liées à la mobilité », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Francony J.-M., « L'éditorialisation des données aux bornes des API : Enjeux et perspectives pour une analyse empirique », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Joux, A., Bassoni, M., « Le journalisme saisi par les Big Data ? Résistances épistémologiques, ruptures économiques et adaptations professionnelles », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Kitchin, R. *The Data Revolution*, Londres : Sage, 2014.

Koch, O., « Les données de la guerre. Big Data et algorithmes à usage militaire », dossier : Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, *Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Lehmans, A., « Les réinventions de la démocratie à l'aune de l'ouverture des données : du discours de la participation aux contraintes de la gouvernance », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Maresca B. « Les Enquêtes de fréquentation des bibliothèques publiques ». *Bulletin des bibliothèques de France (BBF)*, 2006, n 6, p. 14-19. [en ligne] http://bbf.enssib.fr/consulter/bbf-2006-06-0014-003 (consulté le 25 septembre 2018)

Mattelart A., *Histoire de la société de l'information*, Paris : La Découverte, 2001.

Mayer-Schönberger V et Cukier K., *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Boston : Eamon Dolan/Houghton Mifflin Harcourt, 2014.

Miège B., *L'information-communication, objet de connaissance*, Paris : INA Editions, 2004.

Rossi J. et Bigot J.-E., « Traces numériques et recherche scientifique au prisme du droit des données personnelles », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Sire, G., « Web sémantique : les politiques du sens et la rhétorique des données », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Théviot, A., « « Une économie de la promesse » : mythes et croyances pour vendre du Big data électoral », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

Touraine A., *Production de la société*, Paris : Le Seuil, 1973.

Vayre, J.S., « Les machines apprenantes et la (re)production de la société : les enjeux communicationnels de la socialisation algorithmique », dossier : *Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication, Les Enjeux de l'Information et de la Communication*, n° 19/2, 2018.

To pay or not to pay : les musiciens à notoriété locale face à la publicité ciblée sur Facebook

To pay or not to pay: local fame musicians face Facebook targeted advertising

To pay or not to pay: músicos con notoriedad local frente a la publicidad dirigida en Facebook

Article inédit, mis en ligne le 15 novembre 2018.

Caroline Creton

Caroline Creton est doctorante en Sciences de l'Information et de la Communication à l'Université de Rennes 2, membre du laboratoire PREFics et enseignante ATER à l'IUT de Rennes. Sa thèse porte sur l'intégration des TNIC dans les pratiques communicationnelles au sein des scènes locales de musiques actuelles. Elle est encadrée pour ce travail par Anne-France Kogan, Université de Rennes 2 et Inna Lyubareva, IMT Atlantique.
caroline.creton@outlook.fr

Plan de l'article

Résumé

Facebook, plateforme promotionnelle incontournable de la filière musicale, propose aux musiciens un service payant de publicité ciblée. Par la détention de données massives sur les goûts culturels des utilisateurs, la firme laisse entendre qu'un appariement entre musicien et public serait possible grâce au *big data*. L'objet de cet article est d'engager une double réflexion sur la position du musicien au sein de cet espace, les nouvelles tâches qui lui incombent, notamment celles d'analyser, d'exploiter et de choisir les données fournies par la plateforme et sur le rôle de Facebook dans les mutations du musicien en artiste-entrepreneur. Il sera question de montrer en quoi l'offre industrielle de Facebook participe aux mutations qui font du musicien un artiste-entrepreneur.

Mots clés

Facebook, artiste-entrepreneur, *big data*, publicité ciblée, musiciens à notoriété locale, autopromotion, musiques actuelles.

Abstract

Facebook, an essential promotional platform for the music industry, offers musicians a paid service of targeted advertising. By holding massive data on the cultural tastes of users, the platform suggests that a match between musician and public would be possible thanks to big data. The purpose of this article is to engage a double reflection on the position of the musician within this space, the new tasks that are incumbent upon him, especially which of analyzing, exploiting, and choosing the data provided by the platform and on the role of Facebook in the changes of the musician in artist-entrepreneur. It will discuss how the industrial offer of Facebook participates in the changes that make the musician an artist-entrepreneur.

Keywords

Facebook, artist-entrepreneur, big data, targeted advertising, musicians with local fame, self-promotion, popular music.

Resumen

Facebook se ha convertido en la plataforma de promoción imprescindible para la industria de la música. Pues ofrece a los músicos un servicio de publicidad dirigida. Al tener datos masivos sobre los gustos culturales de los usuarios, la plataforma sugiere que un emparejamiento es posible entre el músico y el público gracias al big data. El propósito de este artículo es hacer una doble reflexión sobre la posición del músico dentro de este espacio, las nuevas tareas que le incumben, especialmente la de analizar, explotar y elegir los datos proporcionados por la plataforma y sobre el papel de Facebook en los cambios del músico en artista-emprendedor. Se discutirá cómo la oferta industrial de Facebook participa en los cambios que hacen del músico un artista-emprendedor.

Palabras clave

Facebook, artista-emprendedor, *big data*, publicidad dirigida, músicos con notoriedad local, autopromoción, música actual.

Introduction

« Justice a 1,7 million de fans sur Facebook, mais on ne peut pas leur parler à tous. On ne peut pas sortir du système qui nous est imposé : c'est Facebook qui décide du nombre exact de gens

touchés par une annonce. On est parti de la liberté, de l'échange de fichiers musicaux et de la discussion en direct, pour se retrouver coincé aujourd'hui dans une pseudo liberté »[1]. Ce témoignage questionne la place des artistes au sein des plateformes développées par les acteurs du Web à partir des années 2000 (Beuscart 2007). En proposant des espaces d'échanges et de publication, elles ont attiré un nombre conséquent de musiciens venant y faire leur promotion. Cependant, la massification des usages et le fonctionnement de ces sites ont paradoxalement rendu invisibles un certain nombre de contenus, noyés dans la masse informationnelle, notamment ceux des musiciens à notoriété locale. Ces derniers font le constat amer d'une visibilité possible, mais en réalité limitée, que Facebook leur propose de corriger. À cet effet, la plateforme commercialise un service de publicité ciblée qui consiste à vendre des profils d'utilisateurs, établis grâce aux corrélations recherchées par les algorithmes dans la masse de données collectées (Delort 2015 ; Mosco 2016). En cela, la publicité ciblée est un outil qui relève des possibilités offertes par le *big data.*

La position occupée par les artistes sur les plateformes d'autopublication est celle d'entrepreneur de leur notoriété, comme l'a montré Beuscart avec MySpace (2009), supplanté depuis par Facebook. Cela rejoint les analyses socioéconomiques des industries culturelles qui se sont intéressées à la conversion des artistes en « artistes-entrepreneurs », devant être capables de conjuguer les exigences esthétiques avec des compétences entrepreneuriales (Bouquillion et al. 2015). Dans cette conversion, le succès d'un artiste est imputé à son seul talent et les éléments structurants des industries de la culture, qui y contribuent pour beaucoup, sont omis. Ainsi, pour assurer son développement, l'artiste est appelé à se mettre en scène sur la toile, à consulter les statistiques et potentiellement à acheter des publicités pour renforcer sa visibilité.

Notre article questionne la figure du musicien à notoriété locale, amené à consacrer une partie de son temps à analyser et exploiter les données. En quoi le déploiement des métriques et de la publicité ciblée sur Facebook, issue du *big data*, induisent-ils une révision de la promotion artistique ? Notre hypothèse centrale est que l'offre industrielle de Facebook généralise la position d'artiste-entrepreneur parmi les musiciens à notoriété restreinte, et renforce l'idée que le succès du musicien repose, en partie, sur sa capacité individuelle à utiliser les données massives.

Trois parties composent l'article. La première porte sur les raisons de la pénétration des données massives dans l'activité promotionnelle. La seconde présente les discours portés par les musiciens au sujet du service payant de Facebook. Enfin, la troisième revient sur la manière dont les artistes interprètent les données et endossent l'activité de ciblage publicitaire.

Notre démarche s'inscrit dans les études d'usage qui privilégient une approche compréhensive fondée sur des entretiens qualitatifs. L'article s'appuie sur les résultats de 20 entretiens semi-directifs menés auprès de musiciens des musiques actuelles, sur le territoire nantais de mars à juin 2017 (tab 1). Lors de ces entretiens, l'usage de la publicité ciblée a souvent été évoqué, 13 musiciens avaient réalisé une campagne et d'autres pensaient y recourir. Un questionnaire en ligne a ensuite été envoyé à 182 groupes nantais. Sur 95 réponses collectées — entre septembre et novembre 2017 — comprenant amateurs (47 %) et professionnels (52 %), la moitié avait déjà utilisé la publicité ciblée (50,5 %).

. . . . . .

[1] Pedro Winter, producteur et fondateur d'une maison de disque, dans *Boulevard du Stream* de Sophian Fanen (2017, p.11)

Groupe	Esthétique (description de la page Facebook)	Musique comme activité principale (pour le musicien interrogé)	Publicité ciblée	Communauté Facebook le 11 mai 2018
Koms	Reprise rock	non	non	98
Ma Stol	Progressive Jazzcore	oui	non	893
Rimsa	Tsigane Gypsy Balkan	oui	oui	777
Dolk	Pop - Electro	oui	oui	1 573
Ninon Van	Pop française et chanson délicate	oui	oui	314
Paul Manut	Blues/Rock	oui	non	1 138
Ilia	Chanson pop	oui	oui	1 956
QuéBais	Folk	oui	non	237
Zom Zam	Hip Hop	non	oui	975
Axel Volm	Blues/Folk	oui	oui	711
Eni Rop	Rock/Stoner/Metal	non	oui	1 649
Sam Nibout	Word music/folk	oui	non	807
The Roots League	Reggea	oui	oui	1 109
Surfiler	Indie Electro Pop	non	oui	718
Cobas	Deep-house, deep-techno, house, techno	oui	non	2 771
Orin	Pop française	oui	oui	3 617
Glana	Folk/Rock	oui	oui	2 592
Lodeon	Pop anglaise	oui	oui	4 327
Jain Class	Croon'n'rock'n'roll Garage	non	oui	543
Puka	Musique Tzigane des Balkans	oui	non	2 825

Tableau 1. Groupes interrogés

Pénétration des données dans l'activité promotionnelle des musiciens à notoriété locale

Structuration de la filière musicale

L'arrivée du rock en France, dans les années 1960, génère une amplification progressive de la pratique des musiques populaires (Guibert 2006). La fête de la musique témoigne chaque année de cet élan musical et en 2008 près de 4,2 millions de Français jouaient dans un groupe de musique[2]. Mais parmi ces musiciens, seul un nombre restreint intègre l'industrie musicale.

Les premières analyses socioéconomiques des industries culturelles ont mis en évidence l'existence de plusieurs viviers de main d'œuvre artistique (Huet 1978). Ces viviers renvoient à la position des musiciens vis-à-vis des industries culturelles. Ces dernières, dont les besoins en main-d'œuvre sont limités, supposent une mise en concurrence des artistes souhaitant l'intégrer. Situés en amont de la filière, nombre de musiciens éprouvent un sentiment de fragilité économique et institutionnelle. Cependant, un espace de production coexiste au-delà de l'organisation industrielle de la culture et s'organise dans une économie plus « artisanale », aux logiques marchandes limitées (Miège 2017). Notre enquête s'est déroulée au sein de cet espace de production où se côtoient les pratiques professionnelles et amateurs, dont les acteurs partagent la caractéristique d'assumer de multiples tâches extra-artistiques que recouvre la vie d'un groupe. Selon Garcin, cette situation s'est accentuée avec la crise du disque, laquelle a entraîné la fragilisation économique de l'industrie musicale et la réduction du soutien aux groupes émergents (2015). L'autoproduction est alors devenue la première étape incontournable dans le parcours d'une formation musicale avant de pouvoir bénéficier de l'appui de professionnels de la filière : attaché de presse, tourneur, communicant…

. . . . . .

[2] Enquête du DEPS publiée en 2012, http://www.culturecommunication.gouv.fr/Documentation/Documentation-scientifique-et-technique/Les-pratiques-en-amateur [consulté le 15/10/2017].

Sans accompagnement, les artistes cumulent les fonctions liées à leur projet (Bureau et al. 2009). La promotion artistique est l'une d'elles. Au tournant des années 2000, elle a été recomposée par l'ouverture de nouveaux espaces, propices à la mise en scène de soi. Ces plateformes ont fait tomber le filtre imposé par les *gatekeepers* sur l'information (Cardon 2011) et ont accueilli l'ensemble des musiciens. Cette ouverture de l'espace médiatique a transposé, en ligne, la quête historique de notoriété des musiciens (Beuscart et al. 2015 ; Costantini 2014). Voulant se faire connaître et reconnaître par la filière et le public, ils ont greffé ces nouvelles pratiques à leurs pratiques précédentes de promotion artisanales (Beuscart 2009).

L'usage de ces plateformes d'autopublication a donc été à l'origine de nouvelles tâches ou activités et de manières de faire liées au dispositif technique (Jauréguiberry et Proulx 2011). L'une des dimensions introduites par Facebook, YouTube, etc., dans l'activité promotionnelle est la confrontation des utilisateurs aux données produites sur le Web qui permettent d'évaluer la visibilité et ainsi la notoriété de chaque chose. Cette généralisation des métriques a imprégné l'activité promotionnelle puisqu'elles sont censées refléter l'opinion publique numérique (Bertin et Granier 2015) et par voie de conséquence la notoriété supposée des artistes. Ainsi, les compteurs contribuent à la crédibilité des musiciens auprès des professionnels et servent « *à la fabrication d'un capital réputationnel convertible et monnayable […] sur le marché du travail artistique* » particulièrement instable (Beauvisage et Mellet 2016, p. 94). Le nombre de *likes* constitue, en réalité, un signal vis-à-vis des professionnels.

> « *Plus tu as de likes, plus tu es vu de l'extérieur et plus les programmateurs vont voir. Les programmateurs, ils regardent tout, ils regardent la qualité de ta vidéo, ils regardent la qualité de ton son, ils regardent ce qu'on met, est-ce qu'il a beaucoup joué, où est-ce qu'il a joué, est-ce qu'il est actif, et ils voient aussi ton nombre d'amis, de gens qui te suivent quand ils voient 1700 ils se disent bon bah il y a du monde qui le suit quoi* » Axel Volm, blues/folk, intermittent.

Dans cette logique, pour accroître leur capital réputationnel en ligne, les musiciens se lancent dans diverses actions, manuelles ou payantes, qui témoignent l'importance croissante accordée aux métriques.

> « *Déjà tu invites tous tes amis [à liker ta page]. Tu demandes au maximum de gens qui sont impliqués dans le groupe de faire pareil, les musiciens, management etc. Et puis, quand tu es en concert tu peux demander de liker la page, quand tu fais un post sur ta page et que tu identifies la salle dans laquelle tu vas jouer ou le webzine qui t'a chroniqué ou les personnes qui t'ont accueilli X ou Y, tu peux cliquer sur leur nombre de likes et inviter des gens qui ne sont pas sur ta page : faire le social begger. Mais c'est dur, ça rentre pas, c'est compliqué. Là, on doit être à 1400 en 1 an, c'est pas beaucoup, c'est compliqué à moins de les acheter »* Glana, Folk/Rock, intermittent.

La structuration de l'industrie de la culture incite les musiciens, mis en concurrence, à prendre en charge de la promotion et à prêter attention aux métriques de réputation. Cependant, pour comprendre les usages de ces données, il est nécessaire de les replacer dans le cadre des stratégies capitalistiques des industries de la communication.

Penser l'activité en ligne des musiciens au regard des stratégies industrielles des acteurs de la communication

Les plateformes d'autopublication ont créé des espaces de diffusion regroupant des contenus produits par les industries culturelles et d'autres pouvant être qualifiés d'« *"amateurs" ou "semi-professionnels" […] produits sans contribution financière des acteurs industriels traditionnels* » (Matthews 2015, p. 59). Sans prendre les risques financiers de la production musicale, ces industries de la communication hébergent l'ensemble des contenus culturels à titre gracieux. Elles le font dans le but d'accroître le flux de données et leur possible valorisation économique. Ce modèle

socioéconomique, qualifié de « courtage informationnel » (Rebillard et Smyrnaios 2010 ; Perticoz 2012), se nourrit des échanges culturels des utilisateurs (Bouquillion 2013) afin de développer son industrie publicitaire (Bouquillion et Matthews 2010). Autrement dit, les firmes offrent des services aux musiciens, car cela leur permet en retour de collecter des informations concernant les pratiques culturelles des utilisateurs puis de les valoriser économiquement, potentiellement auprès des musiciens les ayant produites. Ainsi, Facebook entretient une double relation avec les musiciens : d'une part la firme leur fournit un service gratuit, riche en fonctionnalités ; d'autre part elle en fait une cible commerciale pour son service de publicité ciblée, bâti sur les *big data*, que les artistes contribuent à forger en alimentant la plateforme en contenus.

Ces industries, portées par des discours de participation (Matthews 2010), ont montré leur capacité à générer des profits importants (Smyrnaios 2016), grâce à une position dominante dans l'industrie publicitaire en ligne. Dans le cas de Facebook, 97 % des revenus proviennent de cette activité[3] qui s'appuie sur la masse de données captées (Merzeau 2009) au sein de la plateforme et sur des sites partenaires (Mattelard et Vitalis 2014). Cette source de profit structure véritablement le fonctionnement de la plateforme, laquelle a intérêt à écarter des publications pour leur proposer, par la suite, d'en accroître la visibilité en payant (Cardon 2015). Ainsi, les règles de visibilité, établies en toute opacité et en mutation constante, favorisent les contacts réguliers interpersonnels et pénalisent les publications issues des pages publiques, dont celles des musiciens. Cela est orchestré par l'*Edge Rank*, l'algorithme qui organise le fil d'actualité des utilisateurs. Par expérience, les musiciens perçoivent l'endiguement de leur visibilité, opéré par Facebook, au sein même de leur communauté.

> « *Je constate que quand je post un truc sur la page… il y en a peut-être 30 parmi les 800 qui sont abonnés qui voient vraiment la publication. Parce que je vois 800 likes sur ma page mais quand je publie une nouvelle vidéo il y a 30 likes… et puis j'ai pris 40 vues donc en gros je me dis en a 750 qui l'ont pas vu* » The Roots League, reggae, revenu principal issu de la musique.

Ce fonctionnement est mis au grand jour par les différents compteurs qui peuplent la plateforme (*likes*, vues, partages…) et appuient la promotion du service de publicité ciblée. L'extrait suivant et les propos de Pedro Winter, cités en introduction, montrent qu'en suivant les données, les musiciens décèlent les mécanismes de Facebook et certains se sentent contraints de rétribuer la firme.

> « *Quand tu sors un album et que tu as 300 likes sur ta page, quand tu sais qu'en plus même si tu postes un truc ça ne va jamais toucher les 300 personnes, ça va en toucher 20, ils savent que tu es baisé de toute façon, tu es obligé de prendre un sponsoring* » Dolk, électro-pop, intermittent, porteur d'un label.

Cette rétribution, pour accroître la visibilité, prend la forme de publicité ciblée qui consiste en l'achat d'espaces publicitaires sur le fil d'actualité de la propre communauté de l'artiste (ceux qui ont *liké*) ou de nouveaux utilisateurs qu'il est possible de cibler. Par la collecte massive de données, Facebook établit de nombreux critères possibles concernant les caractéristiques sociodémographiques ou socioculturelles comme l'intérêt manifesté dans le domaine musical. Le musicien peut alors sélectionner et coupler à l'infini les critères pour adresser un clip, une annonce de concert à un public jugé cible, d'autant plus que la complexité du traitement algorithmique des *big data*, au cœur de ce service, reste soigneusement cachée. Ainsi, grâce à la mise en forme intuitive des données et pour un montant initial modeste de 10 € environ, la plateforme parvient à adresser ce service aux musiciens à notoriété locale, sans aucune formation marketing.

[3] https://www.lesechos.fr/02/02/2017/lesechos.fr/0211760299478_les-revenus-publicitaires-de-facebook-en-forte-hausse.htm#ge9O3f42rPebAYGd.99 [consulté le 25/02/2018]

Recourir à la publicité ciblée ?

L'usage de ce service questionne néanmoins les mutations à l'œuvre de la figure du musicien en artiste-entrepreneur. Si certains refusent son usage et affirment le rejet des dimensions marchandes des industries de la communication, d'autres l'intègrent. La publicité ciblée serait-elle à l'origine de nouvelles tensions dans la création artistique ?

Mettre à distance les logiques commerciales des industries de la communication

Les musiciens qui rejettent la publicité ciblée ont connaissance de l'outil et savent que d'autres y recourent. Ils affirment leur non-usage comme un choix motivé par leur opposition aux dynamiques industrielles des acteurs du Web et a contrario leur adhésion à une économie « artisanale » de la production musicale.

> « *Tu veux dire, est-ce qu'on paye Facebook pour améliorer notre visibilité ? La réponse est non ! Payer pour Facebook ? Ça fait sacrément chier. Quand on décide de payer des gens, on ne paye pas des multinationales* » Ma Stol, jazz métal, intermittent.

Dans cette même perspective, plusieurs utilisent ou sont sensibles aux logiciels libres de droit et défendent une conception non lucrative du Web.

> « *L'interface du site est sur logiciel libre [...] On peut faire Internet sans pour autant encore filer plein de thunes aux marques qui baignent déjà dans l'argent. J'ai vraiment du mal avec les multinationales [...] Je suis très réticent à l'idée de payer pour sponsoriser du contenu* » Puka, Fanfare balkanique, intermittent.

Au-delà de valeurs politiques, le refus de la publicité relève, dans certaines communautés musicales, d'une logique d'auto-exclusion de ce marché afin de se maintenir dans une frange dite *underground* de l'espace musical.

> « *J'ai surtout une page Facebook. J'ai un profil perso et ma page. C'est mon seul moyen de communiquer avec les gens qui suivent. Si j'avais un site Web ou ce genre de chose, il peut y avoir presque un côté trop marketé, ça peut donner l'image en tout cas... Je sais par exemple que je ne sponsorise pas mes posts Facebook, c'est un choix* » Covas, électro, revenu principal issu de la musique, précédemment Community Manager

Cet ex-Community Manager est le seul à avoir eu une formation spécifique se rapportant à la communication. Connaissant les « ficelles » du métier, il les met pourtant à distance, car elles lui semblent contraires à son ancrage dans l'espace de production musicale où il n'est pas envisageable d'appliquer les logiques commerciales mobilisées dans d'autres secteurs industriels.

Le recours à la publicité ciblée produit de l'incertitude et de la précarité de l'artiste-entrepreneur ?

D'autres musiciens en revanche franchissent le cap et considèrent la publicité ciblée comme un moyen supplémentaire proche des autres formes promotionnelles, plus traditionnelles. L'intégration de la promotion artistique dans des logiques marchandes est alors exprimée de manière lucide.

> « *Je me sers aussi de l'affichage en ville, mais ça commence à mourir les affiches […] aujourd'hui on est vraiment à l'époque des réseaux sociaux, il faut essayer de vivre avec son temps et puis de se servir des outils du moment* », Axel Volm, blues, intermittent, envisage le paiement d'une publicité ciblée.
>
> « *Si tu as vraiment de l'argent tu mets 200 € sur une semaine et c'est matraquage, c'est cool. De toute façon, l'argent c'est le moteur, malheureusement. Si tu sors un truc, la seule différence pour laquelle tu vas être référencé dans les bons magazines, la bonne télé, c'est quel*

> *attaché de presse tu as payé et combien tu l'as payé. Est-ce que tu as payé 2000 € par mois ou est-ce que tu l'as payé 30 000 € par mois ? Si tu en as payé un 30 000 €, qui s'occupe que la presse et un qui s'occupe que la télé, c'est sûr que tu vas être partout* » Glana, folk/rock, intermittent.

Plongés dans l'incertitude de l'artiste-entrepreneur, les musiciens mobilisent le service proposé par Facebook pour valoriser la sortie d'un album ou d'un clip. Les contenus publiés en ligne (clip, teaser…) sont des signaux informationnels destinés aux autres acteurs de la scène (programmateurs, journalistes, public…) afin de faire connaître le travail créatif qui représente, pour un groupe émergent, un investissement conséquent. Cependant, une fois produits, sans l'appui de prescripteurs reconnus, ils en assument seuls la promotion et payent une publicité pour rendre visible cette production[4]. Par ailleurs, les publicités répondent au souci de limiter le risque d'une salle vide, car ne bénéficiant pas d'un contrat de travail protecteur (Guibert, Sagot-Duvauroux 2013 ; Bouquillion et al. 2015), la rémunération des artistes dépend en règle générale de la fréquentation. Aussi, la publicité est vue comme un moyen pour attirer des spectateurs supplémentaires.

> « *C'est quand tous les Irlandais venaient, c'était la grosse date à la salle conventionnée, du coup c'était tout simplement pour faire venir du monde. Là il y avait un réel challenge, on était producteur de notre date, on avait un risque financier* » Mélodie Seas, chanson française, intermittente.

Localement, le musicien peut jouir d'un réseau de prescripteurs en sa faveur et organiser une promotion traditionnelle (flyers, affiches). Mais au-delà de « son territoire », lorsqu'il est en tournée, le manque de relais pour organiser cette promotion motive le recours à la publicité ciblée.

> *« On crée des liens, sponsorisés selon les villes, là on a une date, ça coute 50 balles, pour faire une communication exclusivement sur Saumur.* [Et votre tourneur ne fait pas la com' ?] *Le tourneur n'a pas les bons moyens, moi j'en ai, en fait, j'en n'ai pas non plus, mais, on se démerde »* Dolk, électro-pop, intermittent.

Le travail promotionnel, normalement porté par le tourneur, l'attaché de presse ou la salle de concert, revient aux musiciens. Ainsi, la situation de fragilité vécue au sein de la filière les conduit à user de la publicité ciblée et cela prolonge leur position d'artiste-entrepreneur au sein de l'espace numérique.

Comment endosser le rôle de *data analyst* ?

Parmi les interrogés, seul un musicien dispose d'une formation spécifique en communication et refuse pourtant d'utiliser ce service. D'autres ont pu bénéficier de conseils en communication dispensés par des structures d'accompagnement (Trempolino, les Inouïs du Printemps de Bourges) ou lors de conférences[5]. Ces conseils portent essentiellement sur le nombre de publications par semaine, les bons horaires de publication, la visibilité donnée par Facebook en fonction des types de contenus… Ces différentes « règles » sont d'ailleurs citées par les musiciens qui ont normalisé leur activité promotionnelle sur Facebook. En revanche, ils sont peu accompagnés sur l'analyse des données fournies ou sur la réflexion à mener pour déterminer le public à cibler. Les réponses issues du questionnaire vont dans ce sens, 96 % des musiciens, qui utilisent la publicité ciblée, déclarent la mettre en place, sans appui de professionnels du marketing ou de la communication. Cela transparaît également dans les résultats des entretiens, puisqu'aucun musicien interrogé n'a été

[4] Le premier motif invoqué dans le questionnaire, pour le paiement d'une publicité, est la promotion d'une vidéo (72,9 %)

[5] Se développe le conseil envers les musiciens, à l'instar d'une conférence donnée le 19 septembre 2017, dans une salle de concert, par le Community Manager, intitulée « la Jungle des réseaux sociaux pour un artiste ». Environ 300 personnes étaient présentes, prenant en note les différentes astuces et choses à ne pas faire.

accompagné pour comprendre les données, les analyser ou choisir les critères pertinents pour réaliser la campagne. Dès lors, comment les musiciens à notoriété locale endossent-ils cette activité ?

Suivre et interpréter les données

Le suivi des données préoccupe la plupart des musiciens interrogés. Si certains disent les regarder de loin, elles sont omniprésentes sur la plateforme et impossibles à éviter.

> « *J'ai 1600 likes sur ma page, tous les jours, je vois des likes sur ma page : 1, 2, 5… Cette semaine, mon chiffre est vert, ça veut dire que c'est positif, j'ai douze likes en plus "Whoua, je suis trop forte !", et puis il y a d'autres semaines, le chiffre est rouge parce que j'ai 0 like "personne ne vous a aimé"* » Ilia, chanson pop, intermittente.

Cette avalanche de données est saisie, par certains, comme un outil d'analyse de l'accueil du projet artistique par le public. Le musicien se mue alors en *data analyst* capable d'orienter ses décisions en fonction des données fournies. Il spécule sur le sens des métriques issues des industries de la communication.

> « *Ça me permet de… voir si le concert a bien marché, si ça a pris, s'il y a de l'intérêt…. Vu que je fais des premières parties, ça me montre s'il y a un intérêt par rapport au public que je vais avoir en face de moi. Par exemple quand j'ai fait Stereolux en première partie de XXX dans la salle Maxi devant 1200 personnes, il y a eu très peu de likes après le concert parce que…. c'est un public complètement différent qui se déplace peu au concert [...] Alors que 2 jours après, j'ai joué devant 300 personnes à l'UBU en première partie de YYY, et bah là ça a généré énormément de likes et ça m'a permis de voir que c'est plus intéressant que je fasse des premières parties d'une personne comme YYY que quelqu'un comme XXX* » Orin, pop française, espérant vivre de la musique.

Cette position de *data analyst* au sein de Facebook est renforcée dans l'activité publicitaire qui présuppose une rencontre avec un « vrai » public à la fois enfoui dans les traces collectées par Facebook et retrouvé grâce aux algorithmes. Le musicien doit engager une réflexion sur sa campagne : la durée, le choix du contenu mis en avant, la somme à dépenser et surtout la cible idéale. Pour cela, il peut sélectionner des caractéristiques sociodémographiques comme l'âge, le sexe, le lieu de résidence et les coupler à des critères culturels comme le goût affirmé d'une esthétique musicale, d'un artiste ou d'une salle de concert. Les possibilités sont donc extrêmement vastes et leur couplage sans fin, ce qui permet un niveau élevé d'affinage. Grâce aux statistiques, la firme indique la taille colossale de la communauté concernée par cet intérêt et laisse entrevoir une audience potentielle immense (fig 1).

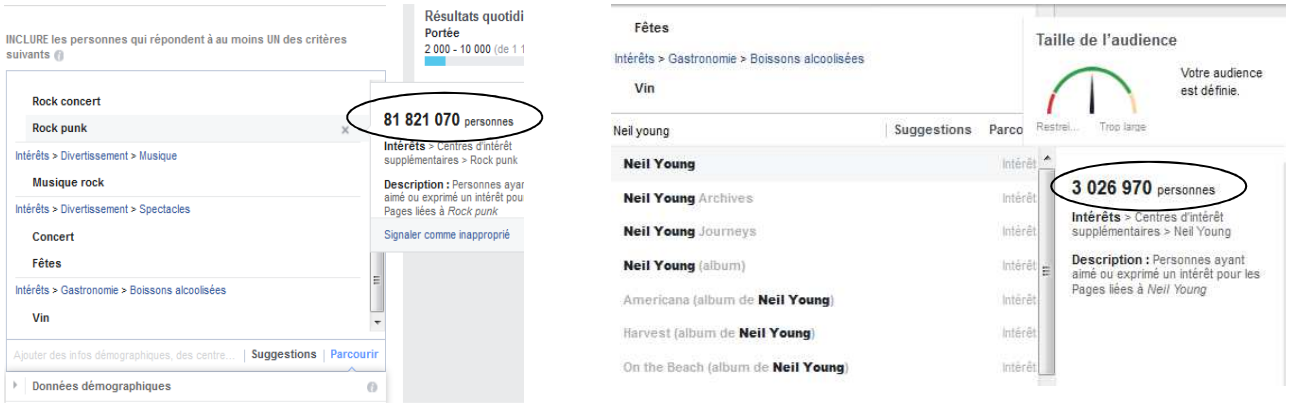


Figure 1. Ciblage et audience estimée, copies d'écran 09/2017

Face à l'ensemble de ces données, les musiciens sont invités à spéculer sur leur public potentiel. Cela reconfigure, en partie, leur activité promotionnelle puisqu'elle élargit leurs domaines de

compétences extra-artistiques aux domaines du marketing et de la communication. Parmi les musiciens interrogés, trois stratégies de ciblage mises en place ont été décelées.

Trois stratégies de ciblage

Ciblage large

Premièrement, une partie des musiciens se gardent d'un ciblage trop précis. Les critères retenus renvoient à des caractéristiques sociodémographiques (sexe, âge, lieu de résidence), ainsi qu'à l'intérêt pour la musique en général ou à une esthétique répandue. Dans cette configuration, la prise de risque et l'efficacité semblent limitées et montrent la difficulté d'opérer des choix plus fins pour circonscrire un public potentiel.

> « *J'ai essayé… enfin je ne suis pas une pro, j'ai essayé de farfouiller un peu. Ah oui ! Je peux créer une audience, je vais cibler les femmes parce que je crois que mon répertoire est plutôt féminin, donc je vais cibler les femmes entre 20 et 65 ans sur un territoire francophone. C'était un test, c'était une audience femme… J'avais essayé un truc plus large, je n'ai pas tenté d'autres trucs parce que ça prend du temps et faut prendre en main l'outil et tout* » Ilia, chanson pop, intermittente.

> « *Je cible une zone géographique… pour les premiers j'avais fait département avec rock et tout, et puis après tu dis "ouais, mais Ben, imaginons tu as un concert à 40 km de chez toi dans ton département t'y vas pas", donc ça sert à rien. Donc quand je joue dans une ville, en gros je fais pas plus de 10 à 15 km en fait […] Pour un enregistrement, un clip, là c'est plus compliqué du coup parce que… c'est des questions que tu te poses comme tu sais pas, c'est pas ton métier, tu demandes quel est le plus efficace. Est-ce que tu fais genre national, ouais pourquoi pas ! Mais en fait c'est ridicule parce que la probabilité de jouer dans leur coin elle est très petite… donc je resserre géographiquement pour essayer de… de toucher, pour avoir un espèce de bouche-à-oreille dans ton environnement, voilà quoi* » Jain Class, rockabilly, espérant vivre de la musique.

Les musiciens expriment leur inexpérience, amateurisme qu'ils assument pleinement. Sans formation préalable ou conseils de professionnels de la filière, ils ajustent le ciblage au fil des campagnes par tâtonnement.

Ciblage pour entrer dans la scène

Deuxièmement, constituer un public cible semble une difficulté moindre pour les musiciens inscrits dans des esthétiques de niche où existent des réseaux établis. Cependant les esthétiques des musiques actuelles n'engagent pas toutes, de la même manière, des communautés de fans. Certaines semblent plus enclines à structurer des communautés au-delà de l'appartenance géographique (Turbé, 2014). Ainsi, l'inscription dans une esthétique accompagnée d'une scène translocale[6] faciliterait le travail de ciblage, le public, qui s'exprime sur ses goûts culturels spécifiques, pouvant être ciblé comme tel. Le critère géographique apparaît moins pertinent pour ce public supposé connecté et prêt à se déplacer.

> « *Dub, reggae, Reggae Roots, Rocksteady, ska, musique afro, soul tout ça quoi. Je mets France entière, dans l'idée je veux développer le projet au niveau national d'ici quelques années puis ensuite internationale donc autant commencer, tout de suite, par le faire connaître au niveau national. Et puis parce qu'en plus Facebook c'est un réseau qui est super large tu peux*

[6] La scène est un concept développé par Bennett et Peterson, ils en distinguent trois formes, la scène locale, la scène translocale et la scène virtuelle. Cette classification, remise en cause par plusieurs chercheurs, introduit une différenciation entre l'effervescence musicale sur le territoire où plusieurs esthétiques peuvent se mêler (scène locale) et une effervescence musicale, plus lâche géographiquement, organisée autour d'une esthétique (scène translocale) (Bennett et Peterson, 2004).

> *toucher quelqu'un à Lille qui a un pote à Nantes… et puis qui va dire "Ah putain, c'est un truc à Nantes" et qui va transférer ça à ses potes sur Nantes* » The Roots League, reggae, vit de la musique.

Ciblage d'une population par d'autres critères

Troisièmement, alors que la plupart des musiciens centrent le ciblage sur l'esthétique musicale ou les caractéristiques sociodémographiques, un musicien déclare mobiliser des critères associés à d'autres intérêts culturels que la musique. Il privilégie des individus au style de vie marqué où la musique apparaît seulement comme l'un élément parmi d'autres.

> « *Je cible les gens qui sont fans de musique, à Nantes, à Paris. Et parce qu'on pense que notre musique s'adresse à un public qui aime la mode, qui aime le côté un petit peu hype, et on en a la preuve on est programmé à la fashion week à Nantes, du coup j'ai ciblé ces gens-là aussi* », Dolk, électro-pop, intermittent, porteur d'un label.

Son usage s'appuie sur des prédictions, réalisées par des intermédiaires professionnels de la filière familiarisés avec le ciblage du public et que le musicien tente de vérifier par la consultation des statistiques fournies par Facebook.

> « *Dans les stats, je vais regarder la situation, la catégorie socioprofessionnelle. Je me suis mis à regarder ça, parce que c'est vrai que depuis le début notre tourneur, nos partenaires en écoutant l'album nous disaient que c'est une musique classe, qui s'adresse à des gens classes* […] *du coup pour vérifier ces dires, je vais regarder, et c'est vrai que c'est souvent des archis, des médecins, des avocats, qui se retrouvent dans notre musique* », Dolk, électro-pop, intermittent.

Par l'utilisation progressive des critères et des statistiques, les musiciens font un apprentissage empirique du web marketing et du travail de *data analyst*. Loin d'un usage homogène, les difficultés sont multiples pour trouver la population jugée « cible ». L'inscription dans une esthétique de niche ou un entourage professionnel semblent pouvoir atténuer cette complexité.

Et finalement ? L'utilité des campagnes en question

Si l'outil de Facebook est construit à partir des *big data*, le caractère prédictif de cette technologie n'est ici pas engagé. Effectivement, les choix des critères reposent largement sur le musicien, à qui est délégué le travail prédiction ce qui rend l'efficacité de la campagne incertaine. Dès lors, comment les musiciens perçoivent-ils le ciblage réalisé et son efficacité à l'issu de la campagne ? À cet effet, Facebook fournit une série de métriques servant à prouver l'utilité de la campagne qui indiquent la visibilité de la publication (portée de la publication, nombre de vues) et l'engagement des récepteurs (partages, commentaires). Les musiciens sont alors confrontés au sens et à l'interprétation de ces données et soulignent l'analyse complexe de l'effet de la publicité sur leur activité réelle.

> « *Je ne peux pas dire* [si c'est utile] *parce que je contrôle pas les gens pour savoir comment ils ont su, pourquoi ils sont venus* », Eni Rop, métal, amateur.

Malgré cette difficulté, les musiciens donnent sens aux métriques au regard des attentes qu'ils ont envers la publicité ciblée. Tout d'abord, certains la conçoivent comme un *amorçage* qui « revient à donner une visibilité initiale à des contenus » (Beauvisage et Mellet, 2016, p. 91). Ils espèrent, en payant, passer le filtre de l'algorithme et initier une circulation sur la toile. Cette idée est reprise sous la métaphore de la boule de neige éclairée par les statistiques de diffusion de Facebook. C'est ici la notoriété qui est en jeu, c'est-à-dire l'évaluation quantitative du nombre de personnes ayant entendu parler du projet. Dans ce cas, la campagne répond aux attentes du musicien qui espère une visibilité enrichie par l'amorçage qu'elle constituerait.

> « *Je pense que c'est utile parce que je vois qu'il y a de nouvelles personnes qui* likent *la page et [...] je vois une publication qui fait quasiment 170* likes *à mon niveau c'est vraiment cool.*

> *Puis je vois beaucoup de partage. Du coup quand quelqu'un partage ça fait forcément boule de neige* » Lodeon, pop anglaise, intermittent.

En second lieu, des musiciens conçoivent la publicité ciblée comme un outil pour gagner en réputation — c'est-à-dire l'émission de jugements qualitatifs sur le projet — et provoquer un *engagement* en ligne, évalué par le nombre de partages et commentaires. Le décalage entre cette attente forte et la visibilité acquise par la publicité se dégage dans l'extrait suivant où les *vues* supplémentaires paraissent insignifiantes puisque non accompagnées d'un engagement.

> « *La promotion n'est pas évidente, trouver un (bon) attaché de presse et pouvoir le payer n'est pas donné à tout le monde, mais c'est la clé… Se dépatouiller avec Facebook et ses incessantes relances pour promouvoir notre publication ça me dépasse. Pour ma part, je paie parfois, j'obtiens 50 likes de plus que ce que j'aurais généré gratuitement, mais les gens ne partagent pas, ni ne vont regarder les vidéos sur YouTube plus d'une fois* » Ilia, chanson pop, intermittente, réponse écrite dans les commentaires du questionnaire en ligne.

Suite à l'usage de la publicité ciblée, la musicienne engage ici une réflexion sur les mécanismes de réputation dans l'espace de production. Ses propos permettent de montrer le rôle essentiel des professionnels de la filière, que les *big data* ne peuvent combler, et de mettre en lumière la fragilité des musiciens situés en amont de la filière culturelle. Cette fragilité, pourtant cachée par le discours de l'artiste-entrepreneur, n'a jamais cessé d'exister.

Conclusion

Les dernières décennies ont été témoins de la conversion de l'artiste en artiste-entrepreneur. Cela est le fruit de la structuration des industries culturelles et politiques publiques (Bouquillion et al. 2015). Cependant, l'offre industrielle des acteurs de la communication, ici de Facebook, prolonge et développe cette figure du musicien-entrepreneur qui doit articuler exigences entrepreneuriales et artistiques. En s'engouffrant dans la promotion en ligne, les musiciens sont contraints par les enjeux socioéconomiques des acteurs industriels de la communication. Si Facebook valorise la participation des citoyens lambda, la plateforme endigue la visibilité de ceux qui en dépendent symboliquement afin d'en tirer une source de revenus. Et cela apparaît sur les nombreux compteurs déployés par la firme qui servent d'argument commercial en faveur de son service publicitaire. Les artistes sont alors confrontés aux multiples données et tentent de leur accorder du sens.

La commercialisation d'un service de publicité ciblée plonge, davantage encore, le musicien dans un rapport aux métriques puisqu'il est amené à choisir les données jugées pertinentes dans la masse que constituent les *big data*. Cet usage est motivé par la fragilité des musiciens au sein de la filière pour qui l'outil paraît atténuer le risque. De ce fait, Facebook brouille la frontière entre l'espace industriel et artisanal de la production musicale, car des logiques marchandes et commerciales s'introduisent dans l'activité de certains musiciens à notoriété locale amateur ou professionnel.

En pénétrant ces espaces culturels, l'industrie publicitaire en ligne contribue au renforcement de la figure du musicien comme artiste-entrepreneur, responsable de son développement devant acquérir des compétences extra-artistiques multiples. Pour endosser le ciblage du public potentiel, les musiciens expérimentent par tâtonnement les possibilités de cet outil, rendu accessible par la firme. S'ils en retirent une visibilité accrue, indiquée par les métriques, cette pratique est l'occasion, pour certains, de questionner le fonctionnement des industries culturelles et de la communication, notamment des mécanismes de notoriété et reviennent sur la nécessité d'un accompagnement par des professionnels de la filière. Cela vient percuter le mythe des possibilités, soi-disant ouvertes par les nouvelles technologies, pour trouver par soi-même les clés de la réputation.

Références bibliographiques

Beauvisage, Thomas et Kevin, Mellet (2016), « Travailleurs du like, faussaires de l'e-réputation, "Like" workers : e-reputation counterfeiters », *Réseaux*, n° 197-198, p.69-108.

Bennett, Andy et Peterson, Richard A. (2004), *Music Scenes: Local, Translocal and Virtual*, Vanderbilt University Press.

Bertin, Erik et Granier, Jean-Maxence (2015), « La société de l'évaluation : nouveaux enjeux de l'âge numérique », *Communication & langages*, n°184, p. 121-146.

Beuscart, Jean-Samuel ; Chauvin, Pierre-Marie ; Jourdain, Anne et Naulin, Sidonie (2015), « La réputation et ses dispositifs », *Terrains & travaux*, n° 26, p. 5‑22.

Beuscart, Jean-Samuel (2007), « Les transformations de l'intermédiation musicale », *Réseaux*, n° 141-142, p. 143‑176.

Beuscart, Jean-Samuel (2009), « Sociabilité en ligne, notoriété virtuelle et carrière artistique », *Réseaux*, n° 152, p. 139-68.

Bouquillion, Philippe (2013), « Socio-économie des industries culturelles et pensée critique : le Web collaboratif au prisme des théories des industries culturelles », *Les Enjeux de l'Information et de la Communication*, [en ligne], Consulté le 21/01/2018, https://lesenjeux.univ-grenoble-alpes.fr/2013-supplement/05Bouquillion/index.html.

Bouquillion, Philippe et Matthews, Jacob Thomas (2010), *Le Web collaboratif : mutations des industries de la culture et de la communication*, Presses universitaires de Grenoble (collection « La communication en plus »).

Bouquillion, Philippe ; Miège, Bernard et Moeglin, Pierre (2015), « Industries du contenu et industries de la communication. Contribution à une déconstruction de la notion de créativité », *Les Enjeux de l'Information et de la Communication*, [en ligne], Consulté le 14/03/2017, https://lesenjeux.univ-grenoble-alpes.fr/2015-supplementB/01-Bouquillon-Miege-Moeglin/index.html

Bureau, Marie-Christine; Perrenoud, Marc ; Shapiro, Roberta (2009), *L'artiste pluriel, démultiplier l'activité pour vivre de son art*, Villeneuve-d'Ascq, France: Presses universitaires du Septentrion (collection « Regard sociologique »)

Cardon, Dominique (2011), « L'ordre du Web », *Médium*, [en ligne], Consulté le 12/03/2018, https://www.cairn.info/revue-medium-2011-4-p-191.htm

Cardon, Dominique (2015), *A quoi rêvent les algorithmes. Nos vies à l'heure des big data*, Paris : Le Seuil (collection « La République des idées »).

Costantini, Stéphane (2014), *Les industries de la musique au prisme des acteurs de l'intermédiation numérique : une analyse des logiques socio-économiques et des pratiques communicationnelles des musiciens*, Thèse de doctorat en Sciences de l'Information et de la Communication, sous la direction de Philippe Bouquillion, Paris 13.

Delort, Pierre (2015), *Le Big Data*, Paris : Presses Universitaires de France (collection « Que sais-je ? »).

Garcin, Pierre (2015), « Devenir musicien dans l'ère numérique », *Sociologie de l'Art*, [en ligne], Consulté le 13/12/2017, http://www.cairn.info/article.php?ID_ARTICLE=SOART_023_0093

Guibert, Gérôme (2006), *La production de la culture : le cas des musiques amplifiées en France*, Paris : Mélanie Seteun / IRMA (collection « Musique et société »).

Guibert, Gérôme et Sagot-Duvauroux, Dominique (2013), *Musiques actuelles : ça part en live. Mutations économiques d'une filière culturelle*, Paris : IRMA (collection « ®evolutic »).

Huet, Armel; Ion, Jacques; Lefebvre, Alain; Miège, Bernard et Peron, René (1978), *Capitalisme et industries culturelles*, Grenoble : Presses universitaires de Grenoble (collection « Actualités - Recherches / Sociologie »).

Jauréguiberry, Francis et Proulx, Serge (2011), *Usages et enjeux des technologies de communication*, Toulouse : éditions Erès (collection « Poche - Société »).

Mattelart, Armand, et Vitalis, André (2014), Le profilage des populations : du livret ouvrier au cybercontrôle, Paris : La Découverte (collection « Cahiers libres »).

Matthews, Jacob Thomas (2015), « Passé, présent et potentiel des plateformes collaboratives. Réflexions sur la production culturelle et les dispositifs d'intermédiation numérique », *Les Enjeux de l'information et de la communication*, [en ligne], Consulté le 20/01/2017, https://lesenjeux.univ-grenoble-alpes.fr/2015/04-Matthews/.

Matthews, Jacob Thomas (2010), « Quelques pistes de réflexion en vue d'une approche critique du Web collaboratif » (p. 329-339), in Millerand, Florence; Proulx, Serge et Rueff, Julien (dir.), *Web social : mutation de la communication*, Québec : Presses de l'Université du Québec.

Merzeau, Louise (2009), « Présence numérique : les médiations de l'identité », *Les Enjeux de l'Information et de la Communication*, [en ligne], Consulté le 01/12/2015, https://lesenjeux.univ-grenoble-alpes.fr/2009/Merzeau

Miège Bernard (2017), *Les industries culturelles et créatives face à l'ordre de l'information et de la communication*, Grenoble : Presses universitaires de Grenoble (collection « Communication en plus »).

Mosco, Vincent (2016), « After the Internet : Cloud Computing, Big Data and the Internet of Things », *Les Enjeux de l'information et de La Communication*, [en ligne], Consulté le 14/12/2017, https://lesenjeux.univ-grenoble-alpes.fr/2016-dossier/09-Mosco/

Perticoz, Lucien (2012), « Les industries culturelles en mutation : des modèles en question », *Revue française des sciences de l'information et de la communication*, [en ligne], Consulté le 11/12/2017, http://journals.openedition.org/rfsic/112

Rebillard, Franck et Smyrnaios, Nikos (2010), « Les infomédiaires, au cœur de la filière de l'information en ligne », *Réseaux*, n° 160/161, p. 163-194.

Smyrnaios, Nikos (2016), « L'effet GAFAM : stratégies et logiques de l'oligopole de l'internet ». *Communication & langages*, [en ligne], Consulté le 31/01/2018, https://www.cairn.info/revue-communication-et-langages1-2016-2-page-61.htm

Turbé, Sophie (2014), « Observer les déplacements dans la construction des scènes locales : Le cas de la musique metal en France », *Cahiers de recherche sociologique*, n° 57, [en ligne], Consulté le 06/10/2016, http://id.erudit.org/iderudit/1035277ar

La production de l'usager par les algorithmes de Netflix

The production of user by Netflix algorithms

La producción del usuario por los algoritmos de Netflix

Article inédit, mis en ligne le 15 novembre 2018.

Gabrielle Silva Mota Drumond

Maître en communication, concentration en médias socionumériques, agente de recherche à l'UQAM, journaliste et gestionnaire de médias sociaux. Elle s'intéresse aux enjeux de la conception des objets technologiques, aux usages des médias sociaux, aux systèmes de recommandation et aux enjeux du traitement de données.

Alexandre Coutant

Professeur au Département de communication sociale et publique de l'UQAM, directeur du Centre de recherche sur la communication et la santé (ComSanté) et responsable de l'axe « Mutations de la sociabilité et de l'agir politique » du Laboratoire sur la communication et le numérique (LabCMO). Ses travaux portent sur l'appropriation des technologies et les enjeux de confiance dans des environnements sociotechniques. Il s'intéresse particulièrement aux contenus en lien avec la circulation de contenus sensibles.

Florence Millerand

Professeure au Département de communication sociale et publique à l'UQAM, titulaire de la Chaire de recherche sur les usages du numérique et les mutations de la communication, co-directrice du Laboratoire sur la communication et le numérique (LABCMO) et membre du Centre interuniversitaire de recherche sur la science et la technologie (CIRST). Ses travaux portent sur les aspects sociotechniques des processus de conception et de développement des technologies ainsi que sur leurs usages.

Plan de l'article

Résumé

Cet article discute de la stratégie de prescription du service de vidéo à la demande Netflix, que nous proposons de qualifier d'« économie de la jouissance ». Il décrit le mécanisme par lequel on stimule un mode de consommation de contenus particulier sur ce service. Ce mécanisme repose sur l'utilisation d'algorithmes qui sont à la base d'un système de recommandation dont l'efficacité est axée sur un parcours de la prescription cohérent. Cet article explore des propositions de méthodes interdisciplinaires pour fournir de l'intelligibilité aux logiques algorithmiques. L'appareillage conceptuel et méthodologique présenté vise à repérer le parcours de la prescription effectué par les concepteurs des dispositifs techniques. À partir d'une approche compréhensive et critique, cet article présente une analyse des choix techniques réalisés par les concepteurs et de la mise en discours des usages attendus qui servent à argumenter et à encourager certaines pratiques chez les usagers.

Mots clés

Conception, usager, systèmes de recommandation, Netflix, algorithmes, prescription, économie de la jouissance.

Abstract

This article discusses the prescription strategy of Netflix, which we propose to call the "enjoyment economy". It describes the mechanism by which one stimulates a particular mode of consumption of contents on this service. This mechanism is based on the use of algorithms that are the basis of a recommendation system whose effectiveness is based on a coherent prescription pathway. This article explores avenues of interdisciplinary methods to provide intelligibility to algorithmic logic. The conceptual and methodological apparatus presented aims at identifying the path of the prescription made by the designers of the technical devices. Based on a comprehensive and critical approach, this article presents an analysis of the technical choices made by the designers and the presentation of the expected uses. This presentation would serve to argue and encourage certain practices among users.

Keywords

Conception, user, software, recommendation systems, Netflix, algorithms.

Resumen

Este artículo analiza la estrategia de prescripción del servicio de video *on demande* de Netflix, que proponemos llamar la "economía del disfrute". Describe el mecanismo por el cual se estimula un modo particular de consumo de contenidos en este servicio. Este mecanismo se basa en el uso de algoritmos que son la base de un sistema de recomendación cuya efectividad se basa en un camino de prescripción coherente. Este artículo explora avenidas de métodos interdisciplinarios para proporcionar inteligibilidad a la lógica algorítmica. El aparato conceptual y metodológico presentado tiene como objetivo identificar el camino de la prescripción hecha por los diseñadores de los dispositivos técnicos. Basado en un enfoque integral y crítico, este artículo presenta un

análisis de las elecciones técnicas realizadas por los diseñadores y la presentación de los usos esperados. Esto último serviría para argumentar y alentar ciertas prácticas entre los usuarios.

Palabras clave

Concepción, usuario, programas, sistemas de recomendación, Netflix, algoritmos.

Introduction

La plateforme de vidéo à la demande Netflix a transformé la consommation de contenus audiovisuels en proposant des offres par abonnement dont l'efficacité du modèle tient à sa capacité à fidéliser les usagers (Boullier, 2009; Citton, 2014; Cochoy, 2004). Pour arriver à ses fins, l'entreprise met en place tout un ensemble de techniques issues de la captologie[1]. Développée à Stanford - avec l'objectif de mobiliser des connaissances en sciences sociales et sciences cognitives pour orienter l'attention des usagers des terminaux numériques (Fogg, 2002) - celle-ci est explicitement appropriée par les industries du numérique dans une perspective de contrôle des activités menées sur leurs services[2]. Dans le cadre de la consommation de contenus culturels, elle aboutirait à la mise en place de ce que nous avons proposé de nommer « économie de la jouissance » (Drumond, 2016). Cette expression sur laquelle nous reviendrons décrit le mécanisme par lequel on stimule une consommation effrénée de contenus. La dynamique a notamment été traitée médiatiquement, à travers la qualification du phénomène du *binge watching* ou écoute en rafale[3]. Elle a constitué le principal ressort pour repositionner l'offre des grands acteurs de la distribution de contenus audiovisuels[4].

Il s'avère que ce nouveau mode de consommation est très populaire chez les usagers de Netflix. Comment expliquer cette popularité? À partir de l'étude du système de recommandation mis en place par la plateforme, initialement appelé *Cinematch*, cet article s'interroge sur le succès de cette offre. Il part de l'hypothèse que ce dernier repose sur une logique de prescription effectuant un parcours cohérent jusqu'aux usagers. Nous tentons donc de mettre à jour comment les attentes du modèle économique de Netflix sont traduites dans un design sociotechnique. Cette question de recherche émerge du constat que la première étape d'un tel design consiste dans l'inscription (Akrich, 1998; Jauréguiberry & Proulx, 2011), dans les algorithmes, d'un usager-idéal correspondant aux attentes du modèle économique de Netflix. La représentation des caractéristiques et actions attendues de l'usager, que l'on retrouve dans le système de recommandation lui-même, émerge d'un processus de configuration de l'usager (Woolgar, 1991) mené par les concepteurs de ce système. Par conséquent, cet article interroge la façon dont Netflix tente de prescrire les usages et faire adapter l'usager à la logique de fonctionnement du logiciel et,

. . . . . .

[1] Standford Persuasive Tech. (2017). *What is Captology?* Consulté 16 mars 2018, à l'adresse http://captology.stanford.edu/about/what-is-captology.html

2 Leslie, I. (2016, octobre). *The scientists who make apps addictive*. Consulté 16 mars 2018, à l'adresse https://www.1843magazine.com/features/the-scientists-who-make-apps-addictive

[3] Jurgensen, J. (2012, juillet 13). *Binge Viewing: TV's Lost Weekends. Wall Street Journal*. Consulté à l'adresse https://www.wsj.com/articles/SB10001424052702303740704577521300806686174

4 Ramachandran, S., & Sharma, A. (2013, septembre 21). *Cable Fights to Feed « Binge » TV Viewers*. Wall Street Journal. Consulté à l'adresse https://www.wsj.com/articles/cable-fights-to-feed-binge-tv-viewers-1379718681

par conséquent, à son modèle d'affaires. Ce niveau de lecture nous apparaît pertinent car la configuration de l'usager peut s'écarter des attentes de l'usager effectif. Elle doit dans tous les cas justifier de la pertinence de ses prescriptions (Hatchuel, 1995; Stenger, 2007) si elle espère être reconnue et inciter à se conformer aux recommandations effectuées (Broudoux, 2007; Quéré, 2001; Verón, 1988). On verra qu'un ensemble de médiateurs sémiotiques est donc mis en place pour présenter les résultats du système au sein de ce dispositif prescriptif (Akrich, 2006a). Cette mise en discours à visée performative constitue effectivement le principal levier pour accompagner l'usager afin qu'il adopte les comportements attendus par la plateforme (Bonaccorsi, 2013; Candel, Jeanne-Perrier, & Souchier, 2012).

Actuellement, Netflix compte 125 millions d'abonnés distribués dans plus de 190 pays dans le monde[5]. Avec un accès illimité à de grandes quantités de contenus comme les séries télé et les films, les abonnés de Netflix visionnent environ 140 millions d'heures de contenus par jour, selon la compagnie. En 2017, le chiffre d'affaires de Netflix avait atteint 11 milliards de dollars américains, ce qui représente une croissance de 33% par rapport à 2016[6]. L'entreprise, fondée en 1997, est aussi un phénomène à Wall Street. En mai 2018, la capitalisation de l'entreprise (153,8 milliards de dollars) a dépassé, pendant quelques heures, celle de Disney (152 milliards de dollars) alors que l'entreprise fondée par Walt Disney représentait jusqu'alors le groupe audiovisuel le plus valorisé au monde[7].

Cet article effectue une analyse détaillée de ce parcours au cours duquel les algorithmes de Netflix sélectionnent et traitent les informations recueillies auprès des usagers pour produire des incitatifs à la consommation, pour tenter d'identifier sur quels leviers ils reposent. Après une rapide présentation de la plateforme, une première partie situe cette recherche au sein du courant des *Critical Algorithm Studies*. Elle spécifie l'approche communicationnelle développée pour effectuer une rétro-ingénierie sociotechnique du dispositif que constitue Netflix. Les enjeux méthodologiques pour développer une telle analyse sont adressés dans une deuxième partie. Les matériaux complémentaires employés sont alors détaillés et mis en correspondance. Une troisième partie analyse le « parcours de la prescription » effectué par Netflix, de son algorithme à sa présentation aux usagers. Une quatrième et dernière partie discute des enjeux soulevés par une telle configuration.

Un cadre théorique pour comprendre les étapes du parcours de la prescription

Une analyse critique des algorithmes sous l'angle communicationnel

Nos activités quotidiennes sont « colonisées » par des technologies numériques auxquelles sont affectées des tâches d'organisation de notre écosystème informationnel (Abiteboul & Dowek, 2017), où l'humain est considéré comme un « nœud » composant les réseaux médiatiques et symboliques qui se modifient constamment (Anderson, 2016). Ceci a donné naissance à un courant

. . . . . .

[5] Netflix Media Center. (2018, mai 31). *À propos de Netflix*. Consulté 31 mai 2018, à l'adresse https://media.netflix.com/fr/about-netflix

[6] AFP. (2018, janvier 22). *Netflix: nouveau gain record d'abonnés au dernier trimestre*. Consulté 31 mai 2018, à l'adresse http://www.journaldemontreal.com/2018/01/22/netflix-nouveau-gain-record-dabonnes-au-dernier-trimestre

[7] CNET France. (2018, mai 25). *Hier, Netflix valait plus que Disney en bourse*. Consulté 31 mai 2018, à l'adresse http://www.cnetfrance.fr/news/hier-netflix-valait-plus-que-disney-en-bourse-39868648.htm

d'études critiques des algorithmes[8] interrogeant les enjeux que soulève cette agentivité toujours plus grande des dispositifs sociotechniques (Boyd & Crawford, 2012; Gillespie, 2016; Kitchin & Dodge, 2011; Steiner, 2013). Nous nous inscrivons dans cette volonté de questionner la pertinence et l'acceptabilité de cette délégation, qui a l'intérêt de soulever au moins trois axes de réflexion. Le premier concerne la non neutralité de ces dispositifs. Si toute une perception objectiviste de l'informatique - fortement appuyée par un discours promotionnel accompagnant ces technologies (Robert, 2016) - tend à assimiler automatisation et neutralité, ces recherches ne cessent d'identifier les choix culturels, politiques, économiques, esthétiques, inscrits dans le code lui-même (Beer, 2016; Cardon, 2015). Une fois cette non neutralité démontrée, ces recherches insistent sur la reproduction, voire l'intensification, des inégalités socioéconomiques traversant nos sociétés (Eubanks, 2017; O'Neil, 2016; Rey, 2017). Enfin, cette mise à jour permet de tenir un discours critique sur les intentions de captation directement inscrites dans ces dispositifs sociotechniques et d'autant plus pernicieuses qu'elles visent à opérer avant toute conscience de choix de la part des usagers (Beer, 2013; Rouvroy & Berns, 2013).

Cette approche critique encourage une lecture communicationnelle de ces dispositifs, attentive aux médiations sur lesquelles reposent leur éventuelle efficacité (Beer, 2013; Rouvroy & Berns, 2013). C'est ainsi que nous nous intéressons au cas de Netflix, à travers l'analyse du parcours de la prescription que ce dispositif suit. Ce parcours peut être reconstitué en deux processus, s'influençant mutuellement et chronologiquement, dont la séparation analytique demeure heuristique. Le premier concerne la composition des algorithmes et leur fonctionnement pour aboutir à des prescriptions. Le second concerne la mise en discours de ces résultats sous un format compréhensible, crédible et désirable pour les usagers. Soulignons que notre approche cherche à comprendre la production et la mise discours de ces recommandations et non leur efficacité. La référence à des enquêtes sur les pratiques culturelles nous permettra cependant d'en discuter le succès en fin d'article.

Dans la lignée de la théorie de l'acteur-réseau (Akrich, Callon, & Latour, 2006), Burrell (2015) qualifie de couche « *hidden* » ou « boîte noire », les algorithmes opérant dans la sélection et la mise en forme des informations auxquelles nous accédons. Il souligne ainsi comment ces espaces opaques condensent des enjeux de pouvoir, qu'un acteur en position de force a fini par rendre non discutables. Décortiquer le dispositif sociotechnique que constituent le système de recommandation *Cinematch* et ses évolutions permet ainsi de mettre à jour les choix effectués pour les interroger à nouveau. La manière dont cette boîte noire produit des prescriptions peut-être décrite en recourant à la notion de script : « il est possible de décrire un objet technique comme un scénario, un script, définissant un espace, des rôles, des règles d'interaction entre les différents acteurs (humains et non-humains) qui viendront incarner ces rôles » (Akrich et al., 2006). En effet, cette lecture permet d'orienter l'analyse vers la définition de la pratique sociale supposée par le dispositif, les acteurs identifiés, les rôles et comportements qu'ils sont censés adopter, les divers éléments contextuels permettant d'exercer cette pratique et d'évaluer son intérêt. Qu'est-ce que les pratiques culturelles pour *Cinematch* ? Dans quel contexte et avec quels acteurs s'exerce-t-elle ? Quels rôles sont prévus et quels objets sont valorisés ? Parmi les éléments composant ce script, nous avons accordé une attention particulière à ceux qui renvoient à l'usager. Woolgar (1991) propose la notion d'usager « configuré » pour insister sur le fait que les concepteurs inscrivent, dans la conception même de l'objet, une représentation des comportements des usagers et de leurs besoins. Pas nécessairement consciente chez les concepteurs, cette inscription aurait la précision d'un profil. Cependant, cet

. . . . . .

[8] Voir la bibliothèque Zotero partagée *Critical Algorithm Studies* pour une recension récente de la littérature sur le sujet : https://www.zotero.org/groups/605005/critical_algorithm_studies

usager configuré ne recoupe pas intégralement l'usager effectif. Il en constitue une représentation réduite, incarnée dans le choix de retenir certaines informations et d'en abandonner d'autres. Ces choix s'avèrent essentiels pour reconstruire l'intentionnalité incarnée dans le dispositif sociotechnique : quel usager idéal produit-on, quels comportements doit-il adopter et pour servir quelles fins des producteurs de ce dispositif ou du dispositif lui-même ?

La reconnaissance à travers la mise en discours des prescriptions

Disposer de contenus à prescrire ne constitue que la première étape du parcours de la prescription. Pour que celle-ci ait une chance d'être reconnue comme valable et digne de confiance, encore faut-il qu'elle corresponde à la grammaire de reconnaissance que lui appliquera l'usager (Coutant & Domenget, 2014). Or, ni la décomposition de ce dernier en ensembles de traces d'activités, ni les différents traitements et mises en correspondance qui aboutiront à une sélection de contenus et à leur classement, ne lui sont directement compréhensibles. Plus encore, la divulgation de la masse d'informations accumulée à son sujet tout autant que des critères de sélection, souvent bien éloignés d'un rapport romantique à la consommation culturelle, risqueraient de plonger ces moteurs de recommandation dans une vallée de l'étrange le rebutant[9].

Une opération de traduction est donc nécessaire pour doter ces prescriptions de suffisamment de gages de confiance pour qu'elles soient acceptées (Quéré, 2001). Nous entendons ici la traduction comme « (l')ensemble des négociations, des intrigues, des actes de persuasion, des calculs, des violences grâce à quoi un acteur ou une force se permet ou se fait attribuer l'autorité de parler ou d'agir au nom d'un autre acteur ou d'une autre force » (Akrich et al., 2006). Il s'agit effectivement de repérer comment l'interface de Netflix incite à déléguer au dispositif le choix des contenus à consommer. C'est ici que la notion de prescription telle que développée par (Hatchuel, 1995) et appliquée au numérique par (Stenger, 2007, 2011) s'avère particulièrement riche. Cette dernière peut effectivement être entendue comme une influence potentielle. Pour qu'elle se réalise, il faut que le prescripteur et le contenu qu'ils proposent soient reconnus comme des aides à la décision. L'intervention doit donc apporter des savoirs qui font défaut à l'usager et faciliter ainsi le processus de consommation. Elle éclaire donc bien comment la mise en discours de Netflix va accumuler les signes la plaçant dans une position de connaissance à l'égard du catalogue disponible et d'experte dans les critères à appliquer pour les évaluer – que ces critères proviennent de ses propres cadres ou de sa connaissance des critères employés par des tiers, notamment les autres usagers. Au-delà des énoncés, la variété des formats énonciatifs encourageant la reconnaissance (Verón, 1988) constitue effectivement une clé essentielle de compréhension des succès variés des moteurs de recommandation selon les contextes (Coutant & Domenget, 2014).

Une méthodologie adaptée à l'analyse du parcours de la prescription

La complexité de l'analyse des réseaux d'acteurs (Akrich, 2006a, 2006b) dans lesquels les algorithmes sont développés apporte des défis méthodologiques inédits[10]. D'une part, ces défis renvoient à la complexité technique des logiciels et aux transformations que ces objets ont subi au fil du temps, nécessitant des connaissances différentes de celles accumulées traditionnellement en

. . . . . .

[9] InternetActu.net. (2012, avril 26). *Les limites du ciblage publicitaire personnalisé*. Consulté 16 mars 2018, à l'adresse http://www.internetactu.net/2012/04/26/les-limites-du-ciblage-publicitaire-personnalise/

[10] Pour une présentation plus exhaustive de ces défis et des manières d'y répondre, voir Drumond (2019, à paraître ; 2016)

sciences sociales. D'autre part, l'absence d'outils méthodologiques spécifiques permettant un repérage minutieux du parcours de la prescription réalisé par les concepteurs invite à faire preuve d'innovation méthodologique.

Un premier enjeu concerne la compréhension du système prescriptif lui-même. Compte tenu des transformations des logiciels et de leur composante algorithmique, les « boîtes noires » peuvent se présenter comme des objets d'études opaques. La compréhension minimale des notions en génie logiciel et langage algorithmique s'avère une nécessité dans ce type de recherche. Cependant, la compréhension de ces mécanismes ne doit pas pour autant détourner de leur analyse critique avec les grilles de lecture plus traditionnelles des sciences sociales. Notre solution face à ces difficultés est passée par la coopération de deux chercheurs spécialistes d'informatique. Ils nous ont accompagné dans l'analyse critique des choix techniques de l'algorithme en traduisant ceux-ci en objectifs compréhensibles pour des non experts. Ceci a permis d'assurer la rigueur de la démarche scientifique et d'éviter les méprises d'information (Drumond, 2019). Le second enjeu consiste à se donner le moyen de suivre la mise en discours des résultats obtenus par ce système prescriptif d'une manière qui soit compréhensible et désirable pour l'usager (Grignon, 2016). La « traduction » des informations techniques en des contenus plus vulgarisés joue effectivement un rôle important dans l'analyse des logiques prescriptives inscrites (Akrich, 2006b) dans les logiciels car elle aboutit à une interface à travers laquelle les usagers vont décider de se conformer ou non aux prescriptions du système.

Notre approche repose sur un ensemble mixte de méthodes compréhensives de recherche. Elle rend possible le repérage et l'analyse de ce parcours de la prescription en deux temps : les choix techniques effectués lors de la conception et la mise en discours des usages prescrits (Akrich, 2006a). D'une part, ces choix techniques effectués au sein de l'environnement de production de l'entreprise Netflix collaborent à la construction d'une logique prescriptive bien précise. D'autre part, les usages configurés par les concepteurs du logiciel de ce service sont soigneusement réaffirmés et argumentés dans le discours employé par l'entreprise. Ce discours peut être visualisé à l'interface du logiciel et identifié dans les instructions données par les concepteurs afin de coordonner les usages (Akrich, 2006b).

Les choix techniques des concepteurs : une étude du système de recommandation et du TechBlog de Netflix

Le processus de repérage du parcours de la prescription s'initie par l'analyse des scripts de deux objets d'études : le système de recommandation, que nous qualifierons de boîte noire, et l'interface présentée aux usagers de Netflix. Pour avoir accès à ce service de vidéo à la demande, les personnes intéressées doivent s'abonner au service à partir de la création d'un compte. Les usagers partageant le même compte peuvent créer des profils qui leur permettront d'avoir accès aux listes de contenus de la plateforme (films, séries, etc.). Ces contenus sont organisés en rangées horizontales sur l'interface. L'usager peut dérouler la page verticalement pour découvrir les différentes rangées de contenus recommandés et horizontalement pour explorer les contenus appartenant à la rangée qui l'intéresse.

Puisque les sections de l'interface ne sont composées que de prescriptions, le système de recommandation représente un élément central chez Netflix. Il peut être considéré comme un espace où une représentation de l'usager, fondée sur une sélection de traces, et des contenus sont mis en relation dans un processus de modélisation (*data mining*). L'emploi de certains critères de classement de contenus au détriment d'autres ainsi que les méthodes de calcul utilisées pour effectuer des appariements entre usagers et contenus potentiellement désirés produiront des effets sur ce qu'on voit à l'écran qu'il s'agit d'expliciter. L'analyse de cet espace de production des prescriptions s'est faite à partir des documents diffusés sur Internet par Netflix entre 2009 et 2015 à

propos de son système de recommandation. Précisément, la solution Bellkor, retenue au terme du concours Netflix Prize[11], et des publications du Tech Blog de l'entreprise ont constitué un premier corpus pour cette étude. La grille d'analyse comportait quatre catégories: représentation de l'usager (attributs associés à la version configurée de l'usager), contenus (attributs associés aux contenus présents sur la plateforme), modélisation des résultats (classement et appariement effectué en fonction des attributs associés à la représentation de l'usager et aux contenus) et résultats (ce que produit le système de recommandation). Nous avons ainsi pu identifier et classer par année les critères utilisés par l'entreprise dans la création des recommandations.

La mise en discours des prescriptions à destination des usagers

Le deuxième corpus analysé a été composé de captures d'écran de l'interface du service et du site web. La justification des choix présentés à l'usager et l'encouragement à certains types d'usages peuvent effectivement être repérés sur l'interface de Netflix et sur le site de l'entreprise. Les usages imaginés par les concepteurs y sont encouragés et argumentés, comme dans les titres des rangées de contenus et les conseils donnés sur le site web de l'entreprise et à travers la fonction permettant aux usagers de regarder des épisodes en boucle. Une analyse sémio-communicationnelle (Bonaccorsi, 2013; Candel et al., 2012; Davallon et al., 2013)[12] a été effectuée afin de saisir ces signes. Deux grilles d'analyse nous ont permis de les identifier et les classer.

Au-delà du repérage et de la description des éléments visuels et textuels ainsi que des icônes de l'interface, ces grilles d'analyse rendent intelligibles la représentation de l'usager énoncée dans le discours de l'entreprise et la façon dont le système prescriptif est justifié. Elles nous ont aussi permis d'identifier les actions permises sur le service et celles suggérées par la plateforme. Par exemple, la rangée « *Continue Watching* » favoriserait la reprise des contenus récemment écoutés et appréciés par les usages, alors que la rangée «*Watched by* (nom d'un personnage des productions originales Netflix) » justifierait auprès des usagers la mise en avant de cette prescription. La rangée « *Because You Watched* (titre d'un contenu déjà visionné par l'usager) » met encore plus en évidence les tentatives de Netflix d'expliquer une partie des critères utilisés par ses algorithmes dans la création des recommandations. Ces éléments font partie d'une mise en scène favorisant la confiance des usagers (Quéré, 2001) dans le système et, par conséquent, dans ses prescriptions. Finalement, nous nous sommes penchés sur les éléments textuels et visuels visant à enseigner l'utilisation du service aux usagers.

L'économie de la jouissance : ce que Netflix enseigne à ses usagers

La trajectoire des transformations du système de recommandation, l'interface du service et le site web mettent en évidence les efforts déployés par Netflix pour optimiser les prescriptions de son logiciel. Ces prescriptions participent à produire à la fois l'usage et l'usager de la plateforme en suivant un parcours qui consiste à créer des représentations des usagers à l'intérieur du système, à leur associer des contenus à suggérer, puis à encourager leur consommation. Ce parcours se base sur quatre éléments centraux : la personnalisation, le classement des contenus et des usagers,

. . . . . .

[11] En 2006, Netflix a lancé le concours mondial Netflix Prize. Ce dernier visait à récolter des propositions qui pouvaient améliorer de 10% l'efficacité de son système de recommandation, appelée *Cinematch* à l'époque.

[12] Conformément à l'approche désormais largement employée (Barats, 2017), consistant à analyser un objet communicationnel (Davallon et al. , 2003) en considérant que le système technique et son concepteur, le dispositif servant d'interface et les usagers sont en situation de communication au sein d'un « théâtre figuratif de la communication » (Boutaud & Berthelot-Guiet, 2013, p. 5).

l'adaptation du logiciel et les feedbacks. Ces éléments sont exposés et expliqués aux usagers par Netflix afin de les inciter à se conformer à son plan d'action. Les modifications du système de recommandation se sont révélées sous différentes formes. La mise en place d'un système de profils en 2013 a contribué à individualiser l'usage du service et, par conséquent, à aiguiller les calculs algorithmiques de son système de recommandation. Ceci a contribué à rendre l'usager plus ou moins conscient des travaux techniques et dans tous les cas acteur du service par son envoi de feedbacks sous la forme de nouvelles données rentrantes. Cette conscience des travaux algorithmiques émerge aussi des recherches de Thoer, Millerand, Vrignaud, Duque, & Gaudet (2015) sur l'appropriation des plateformes de visionnement de contenus. Des entrevues et des focus groupes réalisés auprès de jeunes Québécois montrent que les usagers seraient conscients des travaux de curation des contenus et chercheraient aussi à se conformer à la logique des plateformes afin d'obtenir une meilleure expérience de visionnement.

Pour Netflix, les travaux de personnalisation permettent de mieux cibler les préférences des profils d'usagers et, par conséquent, de leur retourner un meilleur service. Cette personnalisation est rendue visible à plusieurs endroits sur l'interface du service, comme dans les rangées « *Top Picks For* (nom du profil) » et « *My list* » ainsi que sur la page « *Manage Profile* ». Le profil utilisé durant les visionnements est rendu visible dès la connexion au compte. Par ailleurs, la personnalisation est abordée sur le Tech Blog et expliquée sur le site web de l'entreprise. En l'occurrence, on y explique qu'elle calibre le système et rend possible la création des listes de recommandations, qu'elle réduit le temps pour effectuer un choix et qu'elle permet de retenir les usagers sur la plateforme. Ces travaux de personnalisation de la plateforme illustrent la « version configurée » de l'usager effectif. Les calculs algorithmiques sont calibrés par de gros paquets de données rentrant dans le système de recommandation afin d'aligner la « version configurée » sur cet usager effectif.

Le classement des produits et des usagers est ancré dans la conviction que les préférences et les goûts peuvent être mesurés et saisis par les mécanismes de prédiction des logiciels (Beer, 2014). Il s'agit d'identifier des « *patterns* » de comportements et de faire des appariements. Dans le cas de Netflix, son système de recommandation, relevant du « filtrage collaboratif », compare le profil d'usagers à des groupes d'usagers ayant des comportements semblables afin de produire des listes de recommandations. La représentation de l'usager découle ainsi du pouvoir sémantique et de l'intervention politique qu'exercent les concepteurs sur la catégorisation des publics. Ce travail de classement est mis en évidence tant sur le Tech Blog que sur l'interface du service. Ainsi, plusieurs attributs permettant de classer et de hiérarchiser les contenus et les profils d'usagers, ont été introduits dans le système de 2009 à 2015. Par exemple, dans les textes du Tech Blog publiés en 2012, la prédiction des notes apparaissait comme une mesure importante dans les calculs de Netflix. Le système de recommandation du service cherchait en effet à estimer les notes attribuées aux contenus par les usagers, afin de repérer ceux qui étaient susceptibles de recevoir les plus hauts scores. D'autres attributs, telles que la popularité et la diversité des contenus, ont été intégrés au système au fil du temps.

Les travaux de Netflix se concentrent également sur l'adaptation du logiciel aux différents environnements où l'usage se développe. Dans les textes de 2014 extraits du Tech Blog, l'entreprise relate les développements en matière de « *Quality of Experience* » (QoE) et de « *Quality of Streaming* » (QoS). Ceux-ci visent à éviter des « bruits » dans les données et à surveiller la qualité de la connexion et des produits offerts par le service. Concrètement, les données provenant du comportement de visionnement sont comparées aux feedbacks des membres afin de garantir la fiabilité des informations rentrant dans le système de recommandation. Par exemple, les données récoltées durant un visionnement ayant subi une interruption en raison de

problèmes de connexion Internet n'étaient plus prises en compte dans les calculs. Netflix justifie ces mesures par le fait que les données ne doivent résulter que des préférences de l'usager.

Les réponses fournies par l'usager aux recommandations de Netflix alimentent le système avec de nouvelles données qui seront utilisées dans la composition des prochaines pages de recommandation. Les traces d'usage, les notes données aux contenus, la rangée « *My List* », les commentaires laissés par les membres, le bouton de partage de contenus et la fonction « Post-Play » de l'interface participent à la définition de la version configurée de l'usager. La génération de gros volumes de feedbacks est signalée sur le site web de l'entreprise comme étant décisive pour la qualité du service offert. « Plus vous utilisez Netflix, plus le contenu que nous vous suggérons sera pertinent. » (cité dans Drumond, 2016). Sur ce point, la fonction « *Post-Play* » et le visionnement en boucle des contenus qu'elle permet se présente comme l'une des options fournissant les meilleurs indices sur les préférences de l'usager. Il faut préciser que depuis avril 2017, Netflix a remplacé son système de notation basé sur une échelle d'étoiles par un système de pouce levé/pouce baissé. Selon l'entreprise, le volume d'évaluations des contenus aurait grimpé de 200%[13] grâce à cette méthode plus simple et plus intuitive qui inciterait les usagers à fournir plus fréquemment des feedbacks.

Les éléments visualisés dans l'interface et dans le site web de l'entreprise visent à promouvoir le service et, plus particulièrement, à enseigner les « bons » usages. L'entreprise explique ouvertement les travaux techniques effectués sur son logiciel et l'importance des données personnelles dans la création des recommandations. De plus, la pertinence des recommandations est souvent associée aux gros volumes de données récoltées et à la « bonne qualité » de ces données. Dans le « Centre d'aide », par exemple, les textes suggèrent que l'individualisation de l'usage, c'est-à-dire la création des profils pour chaque usager qui partage un compte, produirait des effets directs sur la pertinence des recommandations, puisque le système s'adapterait aux préférences de chaque profil.

La promotion d'une sorte de « conscience », chez les usagers, de l'existence de leur « version configurée », vise à mettre en lumière les prises qu'ils ont sur le système et ainsi à renforcer les liens de confiance entre les acteurs de ce réseau. Selon Netflix, en connaissant mieux la façon dont les prescriptions sont générées, les usagers feraient davantage confiance au service et collaboreraient davantage à son amélioration : « *This not only promotes trust in the system, but encourages members to give feedback that will result in better recommendations.* »[14]

Ces relations de confiance nourrissent ce que nous avons nommé économie de la jouissance : plus on s'amuse en visionnant et en évaluant des contenus sur la plateforme, plus le système de recommandation sera en mesure de proposer des contenus qui rendront l'expérience Netflix encore plus « *amazing* ». Cette économie est axée sur un mouvement en boucle des informations. Autrement dit, elle se base sur l'entrée de gros paquets de données provenant des expériences jouissives des usagers (« *input* »), sur le traitement de ces informations par une boîte noire bien calibrée et sur la production des listes de contenus qui seront visibles sur l'interface (« *output* »). En expliquant le fonctionnement de son service de recommandation, l'entreprise enseigne en même temps la façon dont les usagers peuvent adapter leurs pratiques afin de mieux intervenir dans les propositions du système. Ce discours de transparence soigneusement argumenté sur le Tech Blog,

. . . . . .

[13] Media Center Netflix. (2017, mai 4). *Goodbye Stars, Hello Thumbs*. Consulté 4 juin 2018, à l'adresse https://media.netflix.com/en/company-blog/goodbye-stars-hello-thumbs

[14] Netflix Technology Blog. (2012, avril 6). *Netflix Recommendations: Beyond the 5 stars (Part 1)*. Consulté 2 avril 2018, à l'adresse https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429

l'interface et le site web encourage les usagers à rester connectés sur la plateforme et à produire des réponses « pertinentes » aptes à préciser encore davantage leur version configurée pour, éventuellement, améliorer leur jouissance. En ce sens, les marathons de visionnement, perçus comme l'expression de la satisfaction des usagers envers les contenus consommés, sont interprétés comme des exemples fiables de ce qui fonctionne comme prescription sur la plateforme. Cette pratique très répandue chez les usagers de Netflix (Jenner, 2016) apparaît tout à fait cohérente avec la logique de consommation en excès et de production de données massives qui organise la captation des usagers des services numériques.

Netflix, la jouissance et la politique culturelle

Ces résultats apportent une contribution originale au regard de la littérature critique concernant les algorithmes, d'abord parce qu'ils n'illustrent pas l'opacité traditionnellement reprochée à ces derniers. Au contraire, Netflix explicite son fonctionnement et apporte même un soin particulier à adapter sa pédagogie à ses différents publics : experts en informatiques, prospects et abonnés[15]. Sa boîte noire est traversée de persiennes. La relation proposée se veut transparente, afin que l'usager puisse comprendre par lui-même tout l'avantage qu'il tire en se conformant aux prescriptions de la plateforme, à savoir : une expérience continue de la jouissance soulagée de l'embarras du choix (Cochoy, 1999) et des insatisfactions provoquées par le *régime d'exploration curieuse* qu'Auray (2016) a identifié comme conséquence de contenus culturels pléthoriques en ligne. Ajoutons que la rhétorique de Netflix assume une conception coconstruite de la qualité du service. On s'approche donc davantage de la lecture du rapport de force décrit à propos des sociétés de contrôle (Deleuze, 1992; Hardt & Negri, 2004), où la mise en conformité des usagers repose davantage sur leur acquiescement, à tout le moins passif, plutôt que sur leur manipulation ou leur contrainte. En somme, le meilleur des mondes plutôt qu'Oceania[16].

Si la critique du différentiel entre opacité des systèmes et transparence des usagers ne s'applique pas de la même manière qu'avec d'autres grands acteurs, notamment les GAFAM, la spécificité du secteur dans lequel Netflix opère soulève des enjeux de politique culturelle. Nos résultats montrent la cohérence du parcours de la prescription mis en place par Netflix. Les études sur le divertissement connecté soulignent quant à elles l'efficacité de ces prescriptions auprès des usagers (Thoer, Millerand, Vrignaud, Duque, & Gaudet, 2015). *In fine*, nos résultats permettent de prendre conscience de l'écart entre une économie de la jouissance et les politiques encadrant usuellement les institutions culturelles (Fleury, 2016; George, Brunelle, & Carbasse, 2015). Les critères qualifiant et classant les contenus prolongent en effet une tradition du plébiscite par les publics. Cette logique de la popularité a déjà été identifiée dans les familles d'algorithmes proposées par Cardon (2015). On peut saisir sa pertinence dans le cadre de finalités de divertissement, mais elle reproduit les enjeux de place accordée aux autres dimensions d'une politique culturelle, traditionnellement regroupées sous les catégories d'informer et cultiver. L'analyse de *Cinematch* et du TechBlog souligne la conscience de Netflix des risques d'enfermement dans des bulles homogènes de contenus (Pariser, 2011) et ses tentatives d'empêcher leur formation. En revanche,

. . . . . .

15 Une récente explicitation de la manière dont les images accompagnant les contenus recommandés permet encore de s'en convaincre : https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76

16 Des Lettres. (2013, juillet 10). *Lettre d'Aldous Huxley à George Orwell* : *« Suggérer au peuple d'aimer sa servitude. »*. Consulté 16 mars 2018, à l'adresse http://www.deslettres.fr/lettredaldous-huxley-a-georgesorwell-la-soif-de-pouvoir-peut-etre-tout-aussi-bien-satisfaite-en-suggerant-au-peuple-daimer-sa-servitude/

ses engagements qualitatifs se cantonnent à deux dimensions. La première se concentre sur la qualité de la bande passante, reproduisant la traditionnelle focalisation sur la prestation technique, et dont on voit actuellement qu'elle permet de soutenir un refus des plateformes d'assumer une politique éditoriale (Sire, 2015, 2016). La seconde concerne les notes et traces d'intérêt (visionnement complet notamment), sans tenir compte des éléments composant traditionnellement la complexité des genres culturels et leur évaluation. On retrouve ici, appliqué au cadre culturel, la réduction décrite par Le Deuff (2012) lorsque l'on passe de la diversité des modes d'appréhension permise par les tags à l'uniformité du *like*. L'économie de la jouissance nous paraît ainsi, au-delà de la complexité technique de son fonctionnement, s'inscrire dans une longue histoire de l'appréhension de la culture sous l'angle de son succès populaire, dont on peut émettre la critique lorsqu'elle aboutit à l'exclusion de tout autre critère.

Ce double constat rappelle à quel point l'analyse sociotechnique gagne à se situer toujours au niveau de la pratique sociale et non au seul niveau de l'usage (Jouët, 1993). Le succès de Netflix repose effectivement sur une *domestication croisée* et non sur une domination : on est satisfait de Netflix si on est actif dans sa configuration (Thoer, Millerand, Vrignaud, Duque, & Gaudet, 2015). Cela nous amène à bien le comprendre et le paramétrer. Il s'agit donc d'un usage plutôt maîtrisé. Cependant, cette maîtrise locale ne change rien aux risques d'uniformisation de l'accès aux contenus culturels et de captation permanente de notre attention qui, paradoxalement, se renforcent avec un usage maîtrisé nourrissant une jouissance toujours plus grande.

Conclusion

Basé sur l'étude des stratégies de prescription des usages mises en place par Netflix, cet article se donnait trois objectifs qu'il espère avoir rempli. Le premier concerne l'argumentation qu'il est possible, au-delà d'une forte prétention du discours technicien à s'approprier la question, de fournir une intelligibilité minutieuse des algorithmes, s'inscrivant dans les traditions compréhensives et critiques des sciences sociales, notamment en information et communication. Le second consiste à fournir des propositions de méthodes interdisciplinaires pour rendre intelligibles ces technologies complexes en cumulant regard critique et compréhension fine de leur fonctionnement. La troisième relève de l'encouragement à multiplier les études de cas afin de ne pas s'en tenir à une position générale face à ces technologies. Les enjeux associés à ces systèmes prescriptifs diffèrent fortement d'un contexte à un autre et selon les usagers auxquels ils s'adressent. La délégation d'une certaine agentivité à une autorité algorithmique ne sera pas aussi sensible lorsqu'il s'agit, par exemple, de sélectionner les contenus les plus susceptibles de nous divertir sur un site humoristique ou de décider de l'actualité à donner à voir sur un site d'information. Les voies par lesquelles le parcours de la prescription réussit ou non à se faire reconnaître varient tout autant. Dans une perspective simondonienne, il s'agit par conséquent de fournir les éléments permettant d'identifier comment le rapport coconstruit, transductif, que nous entretenons avec ces objets, en tant qu'individus et collectifs, contribue à faire advenir des milieux individuant ou aliénant (Guchet, 2010; Simondon, 2007, 2012).

Cardon (2015) a déjà indiqué dans ses travaux que les algorithmes pouvaient être compris par des non experts dès lors que l'on met à jour leur « politique ». De la même manière que le fait de connaître la ligne éditoriale d'un titre de presse nous permet de le qualifier sans nécessairement comprendre les multiples étapes par lesquelles un journal est construit, la mise à jour de cette politique rend possible le positionnement face à ces algorithmes sans pour autant maîtriser les traitements auxquels elle donne lieu. La description minutieuse du parcours de la prescription que nous encourageons complétera cette piste pour une meilleure autonomie des usagers face à ces

dispositifs. D'une part, parce qu'elle offre la possibilité de confronter les traces et traitements effectués à la politique de l'algorithme afin d'évaluer leur cohérence. Il devient ainsi possible de distinguer la reconnaissance de la politique, de la reconnaissance des conditions à travers lesquelles elle s'exerce[17]. D'autre part parce qu'elle vise à mettre à jour le vaste système rhétorique entourant ces politiques pour les rendre si évidentes qu'elles ne seraient plus soumises à l'interrogation (Robert, 2016).

Références bibliographiques

Abiteboul, S., & Dowek, G. (2017). *Le temps des algorithmes*. Paris, France: Éditions le Pommier, DL 2017.

Akrich, M. (1998). Les utilisateurs, acteurs de l'innovation. *Éducation permanente*, (134), 79-90. Consulté à l'adresse https://halshs.archives-ouvertes.fr/halshs-00082051/document

Akrich, M. (2006a). La construction d'un système sociotechnique. Esquisse pour une anthropologie des technique. Dans M. Akrich, M. Callon, & B. Latour (Éd.), *Sociologie de la traduction : textes fondateurs* (p. 179-200). Paris: École des mines de Paris.

Akrich, M. (2006b). Les objets techniques et leurs utilisateurs. De la conception à l'action. Dans M. Akrich, M. Callon, & B. Latour (Éd.), *Sociologie de la traduction : textes fondateurs*. Paris: École des mines de Paris.

Akrich, M., Callon, M., & Latour, B. (2006). *Sociologie de la traduction : Textes fondateurs* (1re éd.). Paris: Transvalor - Presses des Mines.

Anderson, C. W. (2016). News Ecosystems. Dans T. Witschge, C. W. Anderson, D. Domingo, & A. Hermida (Éd.), *The Sage Handbook of Digital Journalism*. London, UK: SAGE.

Barats, C. (2017). *Manuel d'analyse du web* (2e éd.). Armand Colin.

Beer, D. (2013). Algorithms: Shaping Tastes and Manipulating the Circulations of Popular Culture. Dans *Popular Culture and New Media: The Politics of Circulation*. Palgrave McMillan.

Beer, D. (2016). How should we do the history of Big Data? *Big Data & Society*, 3(1), 2053951716646135. https://doi.org/10.1177/2053951716646135

Bonaccorsi, J. (2013). Approches sémiologiques du web. Dans *Manuel d'analyse du web en Sciences Humaines et Sociales* (p. 125-146). Paris: Armand Colin.

Boullier, D. (2009). Les industries de l'attention : fidélisation, alerte ou immersion, Abstract. *Réseaux*, (154), 231-246. https://doi.org/10.3917/res.154.0231

[17] Ainsi, en santé, par exemple, une politique visant à mieux personnaliser le traitement d'un patient pourra être reconnue comme louable. Cependant, les traces accumulées et les conditions de leur traitement pour réaliser cette politique pourront être davantage discutées.

Boutaud, J.-J., & Berthelot-Guiet, K. (2013). La vie des signes au sein de la communication : vers une sémiotique communicationnelle. *Revue française des sciences de l'information et de la communication*, 3.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15, 662-679.

Broudoux, E. (2007). *Construction de l'autorité informationnelle sur le web*. Consulté à l'adresse http://archivesic.ccsd.cnrs.fr/sic_00120710/document

Burrell, J. (2015). *How the Machine « Thinks: » Understanding Opacity in Machine Learning Algorithms* (SSRN Scholarly Paper No. ID 2660674). Rochester, NY: Social Science Research Network. Consulté à l'adresse https://papers.ssrn.com/abstract=2660674

Candel, E., Jeanne-Perrier, V., & Souchier, E. (2012). Petites formes, grands desseins. D'une grammaire des énoncés éditoriaux à la standardisation des écritures. Dans *L'économie des écritures sur le web. Volume 1 : traces d'usage dans un corpus de sites de tourisme*, (Vol. 1, p. 135-166). Paris: Hermès-Lavoisier.

Cardon, D. (2015). *A quoi rêvent les algorithmes*. Nos vies à l'heure des algorithmes. Paris: Le Seuil.

Citton, Y. (2014). *L'économie de l'attention*. Paris: La découverte.

Cochoy, F. (1999). " De l'embarras du choix au conditionnement du marché. Vers une socio-économie de la décision ". *Cahiers Internationaux de Sociologie*, 106, 145-173. Consulté à l'adresse https://hal.archives-ouvertes.fr/hal-00178919

Cochoy, F. (Éd.). (2004). *La captation des publics: c'est pour mieux te séduire, mon client...* Toulouse, France: Presses universitaires du Mirail, DL 2004.

Coutant, A., & Domenget, J.-C. (2014). Un cadre épistémologique pour enquêter sur les dispositifs sociotechniques d'information et de communication. Dans H. Bourdeloie & D. Douyère (Éd.), *Méthodes de recherche sur l'information et la communication*. Paris: Mare et Martin. Consulté à l'adresse https://hal.archives-ouvertes.fr/hal-01352927

Davallon, J., Després-Lonnet, M., Jeanneret, Y., Le Marec, J., & Souchier, E. (2013). *Lire, écrire, récrire : Objets, signes et pratiques des médias informatisés*. Paris: Éditions de la Bibliothèque publique d'information. Consulté à l'adresse http://books.openedition.org/bibpompidou/394

Deleuze, G. (1992). Postscript on the Societies of Control. *October*, 3–7. Consulté à l'adresse http://www.jstor.org/stable/778828

Drumond, G. S. M. (2016, novembre). La configuration des usages sur Netflix : le système de recommandation Cinematch et la représentation de l'usager [Mémoire accepté]. Consulté 17 octobre 2017, à l'adresse http://www.archipel.uqam.ca/9253/

Drumond, G. S. M. (2019). Les usagers configurés par les algorithmes : une proposition méthodologique. Dans *Méthodes de recherche en contexte numérique : Enjeux épistémologiques, méthodologiques et éthiques*. Montréal, Canada: Presses de l'Université du Québec.

Eubanks, V. (2017). Automating inequality: how high-tech tools profile, police, and punish the poor (First Edition). New York, NY: St. Martin's Press.

Fleury, L. (2016). Sociologie de la culture et des pratiques culturelles. Paris, France: Armand Colin.

Fogg, B. J. (2002). Persuasive Technology: Using Computers to Change What We Think and Do. *Ubiquity*, 2002 (December). https://doi.org/10.1145/764008.763957

George, É., Brunelle, A.-M., & Carbasse, R. (2015). *Concentration des médias, changements technologiques et pluralisme de l'information*. Québec, Canada: Presses de l'Université Laval.

Gillespie, T. (2016). Algorithm. Dans B. Peters (Éd.), *Digital Keywords: A Vocabulary of Information Society and Culture* (p. 18-30). Princeton, N.J.: Princeton University Press. Consulté à l'adresse http://culturedigitally.org/2016/08/keyword-algorithm/

Grignon, T. (2016). L'expertise communicationnelle au prisme de ses instruments. L'exemple de Google Analytics. *Les Cahiers Du RESIPROC*, 23-47.

Guchet, X. (2010). Pour un humanisme technologique: culture, technique et société dans la philosophie de Gilbert Simondon. Paris, France: Presses universitaires de France.

Hardt, M., & Negri, A. (2004). *Empire*. (D.-A. Canal, Trad.). Paris, France: 10-18.

Hatchuel, A. (1995). Les marchés à prescripteurs. Dans H. Verin & A. Jacob, *L'inscription sociale du marché* (p. 205-225). Paris: L'Harmattan.

Jauréguiberry, F., & Proulx, S. (2011). *Usages et enjeux des technologies de communication*. Toulouse: ERES.

Jenner, M. (2016). Is this TVIV? On Netflix, TVIII and binge-watching. *New Media & Society*, 18(2), 257-273. https://doi.org/10.1177/1461444814541523

Jouët, J. (1993). Pratiques de communication et figures de la médiation. *Réseaux*, 11(60), 99-120. https://doi.org/10.3406/reso.1993.2369

Kitchin, R., & Dodge, M. (2011). *Code/space: software and everyday life*. Cambridge, Mass: MIT Press.

O'Neil, C. (2016). Weapons of math destruction: how big data increases inequality and threatens democracy. New York, Etats-Unis d'Amérique: Crown.

Pariser, E. (2011). The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. Penguin UK.

Quéré, L. (2001). La structure cognitive et normative de la confiance. *Réseaux*, 108(4), 125-152. https://doi.org/10.3917/res.108.0125

Quéré, L. (2005). Les « dispositifs de confiance » dans l'espace public ». *Réseaux*, no 132(4), 185-217. Consulté à l'adresse https://www.cairn.info/revue-reseaux1-2005-4-page-185.htm

Rey, O. (2017). *Quand le monde s'est fait nombre*. Paris, France: Stock.

Robert, P. (2016). *L'impensé numérique*. Tome 1, Des années 1980 aux réseaux sociaux. Paris, France: Éditions des Archives contemporaines.

Rouvroy, A., & Berns, T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation. *Réseaux*, (177), 163-196. https://doi.org/10.3917/res.177.0163

Simondon, G. (2007). L'individuation psychique et collective: à la lumière des notions de forme, information, potentiel et métastabilité. Paris, France: Aubier.

Simondon, G. (2012). *Du mode d'existence des objets techniques*. Paris, France: Aubier.

Sire, G. (2015). Cinq questions auxquelles Google n'aura jamais fini de répondre. *Hermès, La Revue*, (73), 201-208. Consulté à l'adresse https://www.cairn.info/revue-hermes-la-revue-2015-3-p-201.htm

Sire, G. (2016). Le pouvoir normatif de Google. Analyse de l'influence du moteur sur les pratiques des éditeurs. *Communication & langages*, (188), 85-99. https://doi.org/10.4074/S0336150016012059

Steiner, C. (2013). Automate This: How Algorithms Took Over Our Markets, Our Jobs, and the World. Portfolio.

Stenger, T. (2007). Prescription et interactivité dans l'achat en ligne. *Revue française de gestion*, (173), 131-144. https://doi.org/10.3166/rfg.173.131-144

Stenger, T. (2011). La prescription de l'action collective. *Hermès, La Revue*, (59), 127-133. Consulté à l'adresse https://www.cairn.info/revue-hermes-la-revue-2011-1-page-127.htm

Thoer, C., Millerand, F., Vrignaud, C., Duque, N., & Gaudet, J. (2015). « Sur le web, je regarde des vidéos, des séries et des émissions »: catégorisation et sélection des contenus de divertissement visionnés en ligne par les jeunes de 12 à 25 ans. *Comunicazioni Sociali. Tv Genres in the Age of Abundance. Textual Complexity, Technological Change*, Audience Practices., 2. Consulté à l'adresse http://comunicazionisociali.vitaepensiero.it/scheda-articolo_digital/nina-duque-judith-gaudet-florence-millerand/sur-le-web-je-regarde-des-vidaos-des-saries-et-des-amissions-catagorisation-et-salection-des-contenus-de-divertissement-visionnas-en-ligne-par-les-jeunes-de-12-a-25-ans-001200_2015_0002_0191-313597.html

Verón, E. (1988). *La sémiosis sociale: fragments d'une théorie de la discursivité*. Saint-Denis, France : Presses universitaires de Vincennes.

Woolgar, S. (1991). Configuring the user: the case of usability trials. Dans S. Law (Éd.), *A Sociology of Monsters: Essays on Power, Technology and Domination* (p. 57-99). London: Routlege.

Les enjeux de l'information et de la communication Dossier 2018

Production des données, « Production de la société ». Les Big Data et algorithmes au regard des Sciences de l'information et de la communication

« Une économie de la promesse » : mythes et croyances pour vendre du *Big data* électoral

"An economy of the promise": myths and faiths to sell electoral Big data

"Una economía de la promesa ": mitos y creencias para vender de Big data electoral

Article inédit, mis en ligne le 15 novembre 2018.

Anaïs Theviot

Maîtresse de conférences à l'Université Catholique de l'Ouest, rattachée au laboratoire ARENES (UMR 6051), son travail de recherche développe une perspective sociologique du web politique. Sa thèse analysait l'usage du web par les adhérents, ainsi que les stratégies numériques du PS et de l'UMP en période de campagne électorale. L'organisation de colloques a débouché sur la coordination de deux numéros de revue (Sciences de la société et Politiques de communication), consacrées aux modalités de participation et d'engagement sur Internet. Ses travaux portent actuellement sur les enjeux et effets de la professionnalisation de la Big data électorale en France et aux Etats-Unis.

Plan de l'article

Résumé

Cet article propose une analyse sectorielle en portant la focale sur le champ politique et le temps extraordinaire des campagnes électorales. Il s'agit de porter le regard sur les professionnels du politique qui participent à cette production des données et influent sur les manières de faire campagne. En effet, la technicisation des campagnes électorales s'accompagne de la mise en avant de nouvelles expertises afin de vendre un savoir-faire auprès des partis politiques et des candidats. Les « travailleurs de la donnée », inspirés du modèle américain, affirment pouvoir « prédire » les comportements des électeurs grâce aux big data et ainsi agir « scientifiquement » sur les résultats d'une élection. La création rapide d'agences spécialisées dans la gestion des données atteste du succès de cette rhétorique de l'innovation par les data. Cet article analyse ainsi les discours des prestataires en big data électoral qui cherchent à faire croire en l'efficacité de l'usage de données massives pour remporter une élection afin de légitimer leur profession.

Mots clés

Campagne électorale, *Big data*, croyances, travailleurs de la donnée, partis politiques.

Abstract

This article analyzes the political arena and the extraordinary time of election campaigns. It proposes to study the professionals who participate in this production of the data and influence the manners to campaign. Indeed, the technicization of election campaigns comes along with the emphasis of new expertises. The "workers of the data", inspired by the American model, asserts they can "predict" the behavior of the voters thanks to Big data and so to act "scientifically" on the results of an election. The fast creation of agencies, specialized in the management of the data, gives evidence of the success of this rhetoric of the innovation. This article analyzes the speeches of the professionals who try to persuade in the efficiency of the use of massive data to legitimize their profession.

Keywords

Electoral campaign, Big data, political parties, professionals of data.

Resumen

Este artículo propone un análisis sectorial apoyándose el focal en el campo político y el tiempo extraordinario de las campañas electorales. Se trata de apoyarse la mirada en los profesionales de política que participan en esta producción de los datos e influyen en las maneras de hacer campaña. En efecto, la tecnicización de las campañas electorales se acompaña de la puesta por delante de nuevos peritajes con el fin de vender una destreza cerca de los partidos políticos y los candidatos. Los " trabajadores del dato ", inspirados por el modelo americano, afirman poder "predecir" los comportamientos de los electores gracias al « Big data » y así actuar "científicamente" los resultados de una elección. La creación rápida de agencias especializadas en la gestión de los « data » atestigua del éxito de esta retórica de la innovación. Este artículo analiza así los discursos de los prestatarios de « Big data » que procuran hacer creer en la eficacia del uso de datos macizos para llevarse una elección con el fin de legitimar su profesión.

Palabras clave

Campaña electoral, Big data, creencias, trabajadores del dato.

Introduction

En mars 2018, une enquête menée par le *New York Times*[1] et le *Guardian*[2] a dévoilé que la société Cambridge Analytica[3] avait aspiré, à leur insu, les données de 50 millions d'utilisateurs de

. . . . . .

[1] « How Trump Consultants Exploited the Facebook Data of Millions », *New York Times, by Matthew Rosenberg, Nicholas Confessore and Carole Cadwalladr, 17 mars 2018.*

[2] « Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach », *Guardian, Carole Cadwalladr and Emma Graham-Harrison, 17 mars 2018.*

Facebook ; nombre revu à la hausse par le groupe Facebook qui a déclaré en avril que près de 87 millions d'internautes étaient concernés.

L'usage des données est devenu une pierre angulaire des nouvelles stratégies électorales de partis politiques et interroge quant aux stratégies industrielles et marchandes des vendeurs de pronostics et de techniques de mobilisation électorales. Avoir une base de données bien fournie et de qualité est devenu un atout fort pour un candidat afin de cibler sa communication, mais questionne dans le même temps, la protection des données personnelles et les croyances qui entourent le big data.

L'usage politique des bases de données massives est d'origine américaine. D'abord aux États-Unis et en Grande-Bretagne, des programmeurs ont intégré les équipes de campagne ou ont créé des start-ups visant à travailler les données dans un but électoral. Souvent désignés sous le terme de « news applications », les contenus qu'ils conçoivent sont aussi bien des cartes interactives que des infographies, des bases de données interrogeables ou des logiciels proposant des plans de mobilisation. Leur fabrication repose sur une variété de compétences informatiques permettant de collecter, de traiter, de combiner et de visualiser des données — qu'il s'agisse de chiffres, de textes, de photographies ou de contenus audiovisuels disponibles sur des supports numériques. La spécialisation des professionnels de la politique s'accentue avec la nécessité de développer des compétences techniques (Schneier, 1987). En effet, l'usage des données au sein des partis politiques, notamment en période de campagne électorale (Theviot, 2018) demande la maîtrise de terminologies spécifiques et d'outils numériques. Sur son site, le créateur de NationBuilder[4], Jim Gilliam, souligne : « *Le logiciel est riche, mais il faut appliquer des formules mathématiques pour vraiment en tirer parti* ». Un formateur à l'usage des données en politique confie : « *Tout le monde peut se payer NationBuilder, mais il faut une stratégie data pour aller avec.* » (Entretien réalisé avec un formateur sur l'analyse des data, le 12 décembre 2016). Les nouveaux spécialistes des data cherchent à se rendre indispensable aux partis politiques ou candidats, à légitimer leur expertise et « à faire croire » à la nécessité de leurs compétences afin de gagner une élection.

En France, dès 2009, le rapport de Terra Nova[5], se fondant sur quatre-vingt entretiens réalisés avec les principaux acteurs de la campagne américaine de B. Obama de 2008 à Washington, New York et Chicago, avait mis l'accent sur l'importance de la constitution d'une base de données qualifiée pour le Parti Socialiste (PS) français:

« *Leçon n°5 - les bases de données : la rupture orwellienne. (…) Barack Obama a réussi le rêve orwellien de tout candidat américain : ficher l'intégralité du pays. (…) Elle repose sur la technique du micro-targeting : il s'agit de consolider le maximum de bases de données existantes (bases électorales, commerciales, politiques) afin d'obtenir des données individuelles sur tous les électeurs. Ces données sont utilisées pour élaborer des messages personnalisés, notamment pour le porte-à-porte*. »[6]

[suite de la note]

[3] Cambridge Analytica est une entreprise britannique qui prétend pouvoir déduire, à partir des goûts et préférences des internautes, un profil psychologique ainsi que leurs préférences politiques. Une technique qui nécessite une grande quantité de données personnelles.

[4] Fondé en 2009 à Los Angeles par J. Gilliam, NationBuilder se décrit comme un « système d'exploitation de communauté ». Utilisé lors de la campagne de B. Obama en 2012, il a été employé aussi bien par le Labor Party australien que par Amnesty International, AirBnb ou Handicap International.

[5] Créée en mai 2008, par O. Ferrand, *Terra Nova* est un *think tank* français de gauche qui affiche sa position centrale dans l'expertise socialiste (plus de vingt mille abonnés et près de deux mille adhérents déclarés).

[6] « Moderniser la vie politique : innovations américaines, leçons pour la France », Rapport de la mission d'étude de *Terra Nova* sur les techniques de campagne américaines, 2009, p. 13.

En s'inspirant du modèle américain, la campagne pour l'élection présidentielle française de 2012 s'est appuyée sur des bases de données d'électeurs. Pourtant, cette nouvelle technique n'a pas été mise en avant par les équipes de campagne, se méfiant des remarques possibles sur l'usage des bases de données personnelles (Solove, 2008) et préférant porter l'attention médiatique sur des dispositifs plus attrayants : « *un travail sur les bases de données, sur le nombre de mails collectés, sur les taux de clics, sur les transfo en bases de données, etc. Le niveau de couverture de la presse... (…) ce n'est pas très sexy. Pas très sexy à expliquer, à raconter* » (Valerio Motta, directeur du web à Solférino. Entretien du 21 mai 2012). A l'inverse, cinq ans plus tard, en 2017, l'usage des data est mis sur le devant de la scène médiatique, notamment par les candidats qui cherchent à renforcer leur image de modernité. Ainsi, A. Juppé, pour pallier les critiques sur son âge, a mis en scène l'usage que son équipe faisait de NationBuilder lors de la campagne de la primaire de la droite et du centre, à travers les nombreux entretiens que sa directrice du digital a accordé aux journalistes (Theviot, 2016a). Dans le même objectif, J. L. Mélenchon s'est paré des habits de l'innovation, en déclarant au 20h de TF1, lors de son annonce de candidature à l'élection présidentielle de 2017, user de la même plateforme d'analyse de données que B. Sanders.

En analysant le discours des prestataires en *big data* électoral, nous cherchons à interroger cette mise en scène de l'usage des données en politique comme une « nouvelle science électorale » (Pène, 2013) et questionner ainsi les effets sur la communication politique classique, voire les perceptions et reconfigurations des partis politiques. Il s'agit de s'intéresser aux discours portés par ces acteurs et, plus particulièrement, aux discours proposés notamment par les directeurs d'agences spécialisées en big data électoral qui cherchent à « vendre » leurs expertises auprès des candidats en pariant sur des « économies de la promesse » (Bullich, 2016). Cette étude montre que la construction d'une croyance en l'efficacité du big data électoral en politique, orchestrée par ceux qui vendent ces techniques, a conduit à la normalisation du travail sur les bases de données en campagne électorale, à la réduction des résistances à un militantisme technicisé et rationalisé, couplé à la mise en place d'un nouveau marché professionnel en externe.

Ce travail s'appuie sur une soixantaine d'entretiens réalisés avec les professionnels de la communication numérique des équipes de campagne du PS et de l'UMP en 2012 et 2017, ainsi qu'une trentaine d'entretiens effectués avec les prestataires en *big data* électoral en France (LliegyMullerPons, Spallian, Emakina, Fédéravox, Netscouade, Databox, etc.). L'enquête empirique de 2012 s'est réalisée dans le cadre d'un travail doctoral sur les reconfigurations du militantisme et des organisations partisanes, dont la réflexion se voit poursuivie en portant le regard sur les frontières des partis, *via* l'étude des prestataires qui gravitent aux marges de l'institution, dans un jeu permanent du « dedans » / « dehors ». Ce double terrain effectué en 2011-2012, puis en 2016-2017 permet de mettre en lumière les évolutions des pratiques et des perceptions de l'usage de bases de données en politique.

Faire croire au *big data* électoral

Lors des campagnes électorales, la récolte de contacts s'est longtemps effectuée uniquement de manière artisanale lors de meetings, de réunions publiques ; ce qui change en 2017, c'est l'accentuation de la professionnalisation, associée à une prise de conscience des cadres politiques de l'enjeu stratégique. Le numérique n'est ainsi plus appréhendé seulement comme un gadget de communication pour parler aux plus jeunes (comme cela peut-être le cas avec YouTube notamment). Il devient une porte d'entrée centrale pour susciter des votes en faveur de leur candidat. L'idée est de cibler la communication électorale et de rationaliser le militantisme. Plusieurs actions de mobilisation électorale peuvent ainsi être déployées :

- cibler le porte-à-porte dans tel quartier pour le porte-à-porte dont les habitants seraient en majorité des sympathisants de gauche, mais qui s'abstiennent (Talpin, Belkacem, 2014) ;
- envoyer des mails portant sur les propositions du candidat concernant la sécurité à tel groupe d'électeurs qui se dit sensible à cette thématique ;
- concentrer le boitage dans tel canton en fonction des résultats électoraux des élections précédentes, pour leur adresser par exemple un tract destiné aux électeurs du Front National afin de les inciter à voter pour un autre parti.

Cette catégorisation de l'électorat est rendue possible grâce à l'usage d'une base de données qualifiée. Cela permet à la fois de cibler la communication envoyée et de concentrer l'action militante sur des zones du territoire dans un objectif d'efficience et de performativité (Theviot, 2016b).

Pour répondre aux nouveaux besoins d'expertise sur les data, entre 2014 et 2017, en France, de multiples entreprises spécialisées en collecte, gestion et analyse des data ont été créées. Elles s'inspirent de l'exemple américain, caractérisé par un foisonnement d'agences de communication - Blue State Digital, Echoditto, etc. - qui vendent leurs expertises aux candidats ou à des associations spécialisées dans l'*advocacy* (Issenberg, 2013). En France, l'agence LiegeyMullerPons[7] affiche, dès 2013, sur son site internet une « *expertise en matière de recherche et d'analyse de données au service des partis politiques progressistes en France et en Europe* » (Extrait du site internet de leur agence. http://www.liegeymullerpons.fr/). Ce conseil externalisé en communication politique, dans son versant numérique, auprès des hommes politiques ou des partis, se structurent en France depuis 2012: on peut citer la Netscouade, Spintank, Parteja, LiegeyMullerPons, Spallian[8], DigitalBox, Fédéravox ou Emakina. Ces techniques procèdent ainsi à l'entrée de nouveaux acteurs marchands dans le champ de la production des opinions et des élections. Pour les fondateurs de l'agence LiegeyMullerPons, il s'agit de proposer un logiciel performant qui permet de travailler la base de données des candidats et des élus : « *Progrès de la recherche en sciences politiques, Open Data, Big Data, nouvelles technologies : nous proposons aux candidats et aux élus des outils qui révolutionnent leur manière de faire campagne et d'interagir avec les citoyens*. » (Extrait du site internet de leur agence. http://www.liegeymullerpons.fr/) Cette sollicitation « des progrès de la science politique » par des prestataires en stratégie électorale – pour le dire autrement, cette porosité entre science politique et monde politique (Payre et Vanneuville, 2003) - invite à questionner les relations entre science, marché et politique qui se tissent à l'occasion des élections et contribuant ainsi à la formation d'un régime spécifique de production des connaissances scientifiques. Les prestataires s'appuient d'ailleurs sur des références académiques pour donner une légitimité scientifique à leurs discours. Ainsi, les résultats des travaux d'A. Gerber et D. Green (2008), indiquant qu'un électeur sur 14 peut changer d'avis grâce à un porte-à-porte rationalisé par les data (contre un sur 38 au téléphone et un sur 100 000 par mail), sont régulièrement mentionnés

. . . . . .

[7] Les fondateurs de l'agence française LiegeyMullerPons sont en fait ceux qu'on appelle aussi « les Bostoniens » qui étaient chargés d'orchestrer l'opération porte-à-porte dans sa version numérisée lors de la campagne pour l'élection présidentielle française de 2012. A la fin de la campagne, non-ancrés dans des réseaux politiques, ils rencontrent des difficultés pour se faire une place en Ministère. Mal insérés dans le milieu partisan, les trois Bostoniens ont décidé de monter une agence web spécialisée dans le conseil aux partis politiques européens.

[8] Lors de la campagne municipale à Paris, quand l'équipe d'A. Hidalgo utilisait 50+1, celle de N. Kosciusko-Morizet avait recours à « Corto », solution de cartographie intelligente conçue par la société française Spallian.

par ceux qui vendent ces techniques. Terme étendard, les big data apparaissent pourtant plus comme une marque et « *une promesse commerciale avant d'être un concept ou de désigner une réalité de la pratique des sciences* » (Ollion et Boelaert, 2015). En effet, les injonctions au recours au data seraient en partie construites par ces nouveaux experts de la politique, s'inspirant du modèle américain des *campaigners*[9], qui réactivent des mythes électoraux en tentant de prédire les comportements des électeurs.

Ces « travailleurs de la donnée » proposent de nouvelles prestations pour réduire l'incertitude électorale et sont disposés à y croire. Autrement dit, ils cherchent à légitimer « scientifiquement » leurs nouvelles offres d'outils de campagne, en indiquant que cela peut faire la différence dans le résultat final en faisant gagner 1 à 2 points. Pour construire cette croyance en l'efficacité des data, les promoteurs de ces techniques leur attribuent des effets politiques dans de nombreux pays européens, mettant en scène ainsi un mouvement européen où le big data électoral apparaitrait comme la clé de l'acceptation du changement. La responsable du développement de NationBuilder en Europe va même jusqu'à insinuer qu'à présent, avec un travail fouillé sur les bases de données, n'importe qui pourrait devenir Président !

« *Le changement qu'on a vu dans toute l'Europe, c'est que les nouveaux outils permettent à des partis politiques qui n'existaient pas il y a un an, un an plus tard de gagner une élection, comme ce qu'on a vu à Paris. C'est histoire d'En Marche, elle est hallucinante ! Et on a vu la même chose avec Jeremy Corbyn pour le Parti Travailliste en Angleterre, on voit la même chose en Belgique avec le Parti Communiste qui est en train de détruire le Parti Socialiste en Belgique. On l'a vu de nouveau avec le Parti Travailliste en Angleterre avec le nombre de membres qui ont joints, c'est exceptionnellement plus grand. En Angleterre, il y a un nouveau parti qui s'appelle The Women's Equality Party qui est de nouveau un mouvement qui n'existait pas il y a deux ans et en l'espace d'un an, ils sont devenus vraiment un parti politique solide. Donc je pense que toute cette innovation, tout cette technologie, la barrière d'entrée est beaucoup plus basse qu'elle n'a jamais été et du coup, ça permet à qui que ce soit, d'être président de la République, d'avoir un nouveau parti politique.* » (Toni Cowan-Brown Nation, responsable du développement de NationBuilder en Europe. Entretien du 6 juillet 2017).

D'autres arguent même que leurs prestations en big data électoral seraient un moyen de lutter contre l'abstention croissante et se décrivent alors comme des « bienfaiteurs » de la démocratie : « *Nos analyses montrent qu'il y a un impact significatif car dans les zones ciblées, on est à plus d'une vingtaine de points de hausse en participation et presque dix points de hausse en vote Fillon par rapport à ce qui s'est passé à l'échelle nationale* » (Co-fondateur de Fédéravox. Entretien du 16 juin 2017).

Toutefois sur ce marché du big data électoral, certains prestataires qui ont souhaité en sortir ne tiennent pas du tout le même discours. C'est le cas du directeur de l'agence Spallian - société à l'origine du logiciel de big data électoral « Corto » utilisé lors de la campagne des primaires de la droite et du centre - désormais spécialisée dans le domaine de l'analyse de l'information et des études de sécurité, qui insiste sur les potentiels effets négatifs de ce type de pratiques. Ce nouveau marché du big data électoral semble très concurrentiel et les attaques entre prestataires, lors de nos entretiens, concernant les méthodes utilisées, ont été nombreuses.

. . . . . .

[9] Terme anglo-saxon faisant référence aux professionnels américains de la communication politique spécialisés dans les campagnes électorales.

« *On dérape complètement dans ce domaine. Donc on veut surtout pas être associé aux dérapages qui vont avoir lieu, et j'aime pas trop toute les nouvelles tendances, des 50+1*[10] *et compagnie qui sont vraiment là, sur de l'espionnage de réseaux sociaux quoi, c'est pas comme ça qu'on élèvera le débat et c'est très dangereux même sur un plan démocratique parce que ce genre de pratique, ça va, ça va profiter aux extrêmes. Soit quand les extrêmes vont se saisir de ces esquisses, soit quant au contraire ils vont montrer que, ils vont nourrir toujours les discours de paranoïa qui sont leurs fonds de commerce en disant : 'mais regardez les autres élus vous espionnent'. Donc, c'est une pratique qui est assez dangereuse (…) quand je vois une entreprise qui passe des actions de campagne porte à porte a du démarchage commercial, on peut se demander si on utilise pas les mêmes bases de donnée pour envoyer des mailings commerciaux et envoyer un programme de campagne, sauf que le consentement n'a pas été donné pour les mêmes enjeux* » (Directeur de la société Spallian. Entretien du 8 octobre 2017).

Les possibilités de compilation des données privées se multiplient (Rallet et Rochellandet, 2015) et tendent à créer un danger croissant pour la protection des données nominatives. On serait alors dans une nouvelle forme de surveillance qui réintroduit avec le numérique une sorte de « psycho-pouvoir » (Stiegler, 2008). Le packaging et le marketing effectués par les prestataires du big data électoral semble être une stratégie efficace pour « en mettre plein la vue » et pour éluder, dans le même temps, certains questionnements éthiques, liés notamment à la sécurité (et à la revente) des données privées.

Des candidats et des militants disposés à y croire

Dans une élection, chaque voix compte. Or, l'incertitude électorale est aujourd'hui renforcée par la prétendue volatilité de l'électeur (Swyngedouw et al., 2000), représentation enrichie par les médias au sens où elle alimente l'intrigue électorale (Restier-Melleray, 2002). Cette incertitude conditionne l'*illusio* de la campagne (sa croyance constitutive), le jeu électoral (« rien n'est joué ») et permet de mobiliser jusqu'au second tour les énergies militantes. Mais cette incertitude est source d'angoisse pour les candidats. La professionnalisation des acteurs politiques s'accompagne alors d'un recours croissant aux instruments d'analyse des compétitions électorales (Phélippeau, 2002 ; Lefebvre, 2016) : utilisation récurrente de sondages, analyse des propos tenus sur Twitter - pensé comme un outil de sondage permanent -, adoption de techniques de campagne innovantes pour rationaliser l'action militante. L'usage de bases de données qualifiées participe de ce travail de réduction et de contrôle de l'incertitude en mettant en place des algorithmes et des logiciels afin de paramétrer cette prétendue science électorale. Le travail sur les données permet de renouveler l'instrumentation politique des partis et d'enrichir le répertoire d'actions des candidats. La « croyance médiacratique » développée par E. Darras (2008) trouve autour de la question des données numériques et des médias sociaux de nouvelles incarnations. Autrement dit, les candidats et cadres politiques restent les seuls décisionnaires dans l'usage des data (si bien que le porte-à-porte rationnalisé de 2012 n'a failli par voir le jour faute de soutien interne au PS), mais, peu formés aux enjeux médiatiques et numériques, ils développent une croyance diffuse sur l'impact de ces outils.

. . . . . .

[10] 50+1 est le logiciel de big data électoral vendu par l'entreprise LliegyMullerPons qui agrège les données démographiques de l'INSEE et celles issues des précédents scrutins électoraux. Il est décrit ainsi sur leur site internet : « *50 + 1 combine toutes les données disponibles sur des dizaines de milliers de territoires avec un modèle prédictif afin de comprendre les enjeux d'un territoire et d'analyser l'opinion publique locale* ».

Si, peu de militants socialistes adhèrent au modèle du « parti-entreprise » (Gaxie, 1973) ou au « *business firm model* » (Hopkin et Paolucci, 1999) l'objectif final de conquête du pouvoir (tout comme celui de ne pas faire faillite pour filer la métaphore de l'entreprise) impose une ligne conductrice à l'échelon national : user de tous les moyens légaux qui prétendent avoir un poids sur les résultats électoraux. L'usage du big data électoral fait partie de la boite à outils pour potentiellement améliorer l'efficacité des campagnes électorales. Si en 2012, les techniques de rationalisation de l'activité militante ont été mal vécues par certains militants (notamment socialistes (Theviot, 2014)) ; en 2017, pour l'élection présidentielle, elles apparaissent comme entrées dans la norme pour conquérir le pouvoir. Certains militants attribuent en effet la victoire de F. Hollande au porte-à-porte rationalisé par les data. Cette perception du rôle primordial des bases de données lors de l'élection de 2012 a dissipé les résistances des adhérents du PS qui y voyaient une technicisation du militantisme, le déshumanisant. En l'espace de cinq ans, l'usage des données massives pour rationaliser l'activité militante ne fait plus l'objet des mêmes résistances en interne. Les adhérents ont intégré ces nouveaux outils de campagne dans leur répertoire d'action et ont bénéficié en 2017 de l'apprentissage de la campagne de 2012 ce qui a limité les tâtonnements et le besoin de formations internes. L'ancien responsable du porte-à-porte ciblé par les data de F. Hollande reconnaît le changement entre 2012 et 2017 dans la perception et l'usage des bases de données, lié à l'apprentissage, la normalisation, un certain effet de génération et la croyance dans le « pouvoir » de ces techniques :

« *Ça a incroyablement changé. Ils nous ont très clairement dit, dans l'équipe Hamon : « non, mais vous inquiétez pas ! ». Nos interlocuteurs dans l'équipe Hamon, ce sont les gens que l'on a formé en 2012. Sauf que maintenant, c'est eux qui dirigent la campagne, donc ils connaissent, ils ont déjà fait tout le truc avec nous. Ils savent très bien comment ça marche, tout ce qu'il y a derrière. On a un peu formé la génération suivante, si on compte en cycle de campagne électorale, quoi. (…) Pareil, pour l'équipe En marche, on retrouve des gens qu'on avait rencontrés en 2012. Ou si on ne les connaissait pas, c'est des gens qui ont trente ans, ils travaillent dans des entreprises. En une heure de formation à distance, ils ont compris comment le logiciel fonctionne et c'est bon. En 2012, on contribuait nous-mêmes au changement de pratiques et aux changements de cultures. En 2017, le changement de pratiques est en route, est en train d'avoir lieu, donc on n'a plus les mêmes résistances* » (Arthur Muller, co-fondateur de l'agence LiegeyMullerPons. Entretien du 13 avril 2017).

Les « bonnes » pratiques militantes semblent s'être largement diffusées. Le porte-à-porte rationalisé et modernisé par les data s'est ainsi développé, même à l'échelle locale lors des municipales de 2014 ou lors des législatives de 2017.

Toutefois, les structures partisanes traditionnelles restent emprisonnées dans des schémas classiques et rencontrent encore des difficultés à impulser un changement global pour incorporer en interne une réelle stratégie numérique :

« Il n'y pas de choix d'organisation de la manière de mener une campagne qui sont globaux et permettent une impulsion et un vrai changement de paradigme du parti (…) Repartir de zéro, ça fait peur à tout le monde, ça parait impossible. Alors que c'est de ça qu'on a besoin pour avoir une vraie stratégie digitale, CRM et d'organisation de campagne. On sait pas comment faire dans un vieux parti. Repartir de zéro, c'est plus facile pour En Marche. Pour le PS, ça parait impossible. Alors qu'on pourrait se dire que le PS a une connaissance du terrain qui va lui permettre de mapper, de cartographier, d'organiser mais en fait pas du tout. Il y a une dette technologique dans les grandes organisations. Il y a une dépendance culturelle très très forte qui empêche beaucoup l'innovation, c'est un peu la même chose dans les grandes entreprises. C'est pas un hasard si En Marche arrive en 2016 avec une organisation de campagne hyper souple, hyper agile. » (Chargé

de conseil à la Netscouade sur NationBuilder lors des municipales françaises en 2014. Entretien du 2 août 2017).

En revanche, à l'échelle des primaires en 2016, cela a été plus facile de sortir des façons de faire traditionnelles car la structure partisane était moins prégnante que lors de l'élection présidentielle de 2017. Ainsi, certains candidats tels que A. Juppé, N. Kosciusko-Morizet et B. Le Maire ont fait le pari de s'appuyer sur le jeu de leurs données pour susciter la création de nouveaux groupes. Cette « stratégie du foisonnement » constitue une rupture dans l'organisation classique des partis où la notion de contrôle est centrale. Comme le souligne la directrice de la campagne numérique d'A. Juppé, des animateurs de réseaux ont été recrutés par simple inscription sur le site de campagne : « *N'importe qui peut demander à être animateur, n'importe où, même s'il est dans la même rue que quelqu'un qui en a déjà créé un, dans le même quartier, dans la même ville* » (Eve Zuckerman, directrice du pôle digital de la campagne d'A. Juppé pour la primaire de la droite et du centre de novembre 2016. Entretien du 3 mars 2016). Le risque de confusion est clairement assumé. Le modèle d'inspiration n'est pas partisan. Pour réguler les possibles discordances entre cercles locaux, l'équipe de campagne a fait le choix de s'appuyer sur un réseau de référents – constitué à 30 % d'animateurs qui n'ont jamais fait de politique – pour gérer la coordination et l'organisation du maillage territorial des candidats. Il faut dire que NationBuilder permet de se concentrer sur les sympathisants, en repérant par exemple si un internaute a déjà fait un don ou s'il a été sensible sur Facebook à un argument de campagne. L'objectif est alors de renforcer les liens grâce à une boîte à outils très complète : administration de sites web, création de mails optimisés via un système de test A/B (deux versions différentes du message sont envoyées à deux panels, et la plus efficace est choisie en fonction du taux d'ouverture, du taux de clic…), envoi de SMS personnalisés, gestion des dons en ligne, etc.

Avec la campagne d'E. Macron en 2017, a été franchi un cap dans l'adaptation au décloisonnement des frontières classiques du militantisme, entre adhérent et sympathisant : « *En Marche et le PS, c'est la start up versus la grosse entreprise qui a dû mal à bouger et à sortir du militantisme à papa* » (Chargé de conseil à la Netscouade sur NationBuilder lors des municipales françaises en 2014. Entretien du 2 août 2017). Ce jeune parti n'a pas rencontré les mêmes difficultés internes pour intégrer de nouveaux adhérents, que celles vécues en 2012 lors de l'opération porte-à-porte par le PS. La jeunesse de ce nouveau parti lui évite d'être tiraillé par des enjeux de courants et des pesanteurs dues à une longue histoire militante. En amont de sa campagne électorale, la « Grande Marche » orchestrée par LiegeyMullerPons a permis à l'équipe d'E. Macron de récolter plus de 25 000 questionnaires. Les équipes ont alors réalisé une prospection ciblée sur les personnes indécises ou sympathisants potentiels. Tout au long de la campagne, en accordant une place centrale au recrutement de volontaires par Internet, le mouvement En Marche a su gonfler et déployer, au plus près du terrain, les idées et les programmes. Le recours aux data a permis alors de proposer toute une gamme d'engagements, allant d'une implication ponctuelle (comme retweeter les messages de candidats) à une action plus engageante, comme aller faire du porte-à-porte. En fonction du profil de l'internaute, il lui a été proposé des actions spécifiques. En Marche a ainsi pu bénéficier d'une main d'œuvre militante croissante, composé « *de membres, d'enthousiastes ou de simples curieux* »[11], lui permettant de couvrir le terrain hors ligne et en ligne. Sa communication numérique était d'ailleurs fortement axée sur les sympathisants : par exemple, dans sa newsletter, il

. . . . . .

[11] Extrait du site internet En Marche : « *Osez nous rejoindre ! Vous êtes curieux/se et indécis/e, vous êtes membre d'En Marche mais pas véritablement sûr/e encore : venez discutez avec des gens comme vous et échanger pour vous aider à y voir plus clair.* » (Janvier 2017) ; https://en-marche.fr/evenements/f1df44b7-d4db-510e-9f28-e62afd38bb03/rencontre-pour-les-curieux-et-les-enthousiastes

était systématiquement proposé des argumentaires simples pour convaincre l'entourage, avec un ciblage par thématique.

Le recours au data est d'ailleurs mis en avant par les travailleurs de la donnée comme moyen de soulager les acteurs des équipes de campagne de leur difficulté à mener une mobilisation électorale efficace, faute de militants et d'ancrage territorial : *« Il y a une crise du militantisme en France. Il y a plus de militants. On passe à autre chose. Le jeu maintenant, c'est d'arriver à parler aux sympathisants. On travaille beaucoup avec les équipes de campagne dans nos formations pour leur dire que leur message est compréhensible que par les adhérents : 'Un sympathisant ne connaît pas votre jargon' ou 'là, vous êtes trop direct'. Il faut les aider dans cette transition. Ils ont des difficultés à utiliser un langage société civile. L'outil leur permet déjà de communiquer avec les sympathisants »* (Fondateur de DigitalBox. Entretien du 1er juin 2017).

Conclusion

Lors de l'élection présidentielle française de 2017, les communicants politiques numériques se sont inspirés des techniques de campagne américaines et ont fait du big data électoral un argument politique pour afficher une modernisation des manières de communiquer et revigorer ainsi l'image vieillissante des partis politiques. Entre 2012 et 2017, la perception de l'usage des bases de donnée a largement été modifiée auprès non seulement des cadres politiques qui y voient désormais un atout stratégique (et non un simple outil de communication), mais aussi des adhérents qui ne résistent plus de manière massive à cette technicisation du militantisme car elle apparaît comme un moyen de gagner l'élection. Cette croyance en l'efficacité des data a été construite par ceux qui vendent ces dispositifs et qui ont donc tout intérêt à y croire ou du moins à y faire croire. L'importation progressive de ces techniques américaines tient au fait que ceux qui les portaient en 2012 étaient en partie dominés du fait de leur jeunesse et de leur position dans les échelles d'autorité du parti, « *ces jeunes qui boivent des bières et jouent avec leurs gadgets* » (Lionel, rédacteur en chef du site *françoishollande.fr/webradio* et responsable du pôle Contenu au sein de l'équipe numérique de F. Hollande. Entretien du 2 janvier 2013). En 2017, les jeunes ont grandi, ont monté leurs entreprises et ont réussi à faire croire aux cadres politiques du bien-fondé de leurs ambitions.

Références bibliographiques

Bullich, Vincent (2016), « Big Data : stratégies industrielles et économie de la promesse » (p. 41-71), in Pilati A. (dir.), *La comunicazione multipla. Media, piattaforme digitali*, Over the Top, Big Data, Rome : Magna Carta Edizioni

Darras, Éric (2008) « La coproduction des grands hommes. Remarques sur les métamorphoses du regard politique », *Le Temps des médias*, vol. 10, n°1, p. 82-101.

Gaxie, Daniel (1973), *Les professionnels de la politique*, Paris :PUF.

Gerber, Alan et Green, Donald (2008), *Get Out the Vote: How to Increase Voter Turnout*, Washington D.C. : Brookings Institution Press.

Hopkin, Jonathan et Paolucci, Caterina « The Business Firm Model of Party Organisations : Cases From Spain and Italy », *European Journal of Political Research*, vol. 35, n°3, p. 307-339.

Issenberg, Sasha (2013), *The Victory Lab: The Secret Science of Winning Campaigns*, Reprint edition, New York : Broadway Books.

Lefebvre, Rémi (2016) « La modernisation du porte-à-porte au Parti socialiste. Réinvention d'un répertoire de campagne et inerties militantes », *Politix*, vol. 113, n° 1, p. 91-115.

Mabi, Clément et Theviot, Anaïs (2014), « La rénovation par le web ? Dispositifs numériques et évolution du militantisme au PS », *Participations*, n°8, p. 97-126.

Ollion, Etienne et Boelaert, Julien (2015) « Au-delà des big data », *Sociologie*, vol. 3, n°6, [En ligne], Consulté le 12 septembre 2016, http://sociologie.revues.org/2613

Payre, Renaud et Vanneuville Rachel (2003), « » Les habits savants du politique » Des mises en forme savante du politique à la formation de sciences de gouvernement », *Revue française de science politique*, vol. 53, n°2, p. 195-200.

Pène, Clémence (2013), « La nouvelle "science électorale" américaine », *Politique étrangère*, n°2, p. 127-139.

Phélippeau, Eric (2002), *L'invention de l'homme politique moderne. Mackau, l'Orne et la République*, Paris : Belin.

Rallet, Alain et Rochelandet, Fabrice (2015), Dossier « Protéger la vie privée à l'ère numérique », *Réseaux*, n°189.

Restier-Melleray, Christiane (2002), *Que sont devenues nos campagnes électorales ?*, Pessac : Presses universitaires de Bordeaux.

Schneier, Edward (1987), "Is politics a profession? A new school says yes", *PS: Political Science and Politics*, vol. 20, n°4, p. 889-895.

Solove, Daniel (2008), Understanding privacy, Cambridge : Harvard University Press.

Stiegler, Bernard (2008), *Economie de l'hypermatériel et du psychopouvoir*, Paris : Mille et une nuits.

Swyngedouw, Marc, Boy, Daniel et Mayer, Nonna (2000), « Mesure de la volatilité électorale en France (1993-1997) », *Revue française de science politique*, vol. 50, n°3, p. 489-514.

Talpin, Julien et Belkacem, Romain (2014), « Frapper aux portes pour gagner les élections ? Ethnographie de la campagne présidentielle socialiste dans deux villes du Nord de la France », *Politix*, n°105, p. 185-211.

Theviot, Anaïs (2016a), « Les primaires : terrain d'expérimentation de l'innovation politique ? Le cas de la campagne d'A. Juppé en 2016 : une mobilisation « scientifique » orchestrée par les data » (p. 213-238), in Lefebvre R. et Treille E. (dir.), *Les primaires ouvertes en France*, Rennes : PUR.

Theviot, Anaïs (2016b) « Les data : nouveau trésor des partis politiques ? Croyances, constitutions et usages comparés des données numériques au Parti Socialiste et à l'Union pour un Mouvement Populaire », *Politiques de communication*, n°6, p.137-166.

Theviot, Anaïs (2018), *Faire campagne sur Internet*, Villeneuve d'Ascq : Presse Universitaires de Septentrion.

Les données au service de la connaissance des usages en ligne : l'exemple de l'analyse des logs de Gallica

Data serving the understanding of online uses: the example of Gallica's log analysis

Datos que sirven para el conocimiento de los usos en línea: el ejemplo del análisis de registro de Gallica

Article inédit, mis en ligne le 15 novembre 2018.

Philippe Chevallier

Philippe Chevallier est adjoint au responsable de la coordination de la recherche à la Bibliothèque nationale de France. Docteur en philosophie, il est l'un des fondateurs du « Bibli-Lab », partenariat de recherche entre la Bibliothèque nationale de France et Télécom ParisTech sur les usages du patrimoine numérique des bibliothèques. Il a collaboré avec Anne Monjaret à l'édition de l'ouvrage collectif dirigé par Mélanie Roustan, La recherche dans les institutions patrimoniales : sources matérielles et ressources numériques (Presses de l'Enssib, 2016).

Plan de l'article

Résumé

Connaître les usages d'une bibliothèque numérique comme Gallica nécessite de renouveler les dispositifs d'enquête traditionnellement utilisés par les bibliothèques, en explorant de manière semi-automatisée les données des serveurs. Cette exploration recourt à des modèles mathématiques qui rendent plus difficile le dialogue entre les chercheurs et les professionnels des bibliothèques. Un projet de recherche conduit par la Bibliothèque nationale de France et Télécom ParisTech sur les logs de Gallica témoigne de la possibilité d'inscrire la fouille de données dans un dialogue où les chercheurs et les professionnels s'efforcent de s'éclairer mutuellement sur les décisions à prendre pour conduire une analyse pertinente. Il met également en lumière l'importance de croiser les méthodes scientifiques pour comprendre les usages en ligne, aucune de celles-ci ne pouvant prétendre se suffire à elle-même.

Mots clés

Bibliothèque numérique ; Usage ; Web ; Patrimoine ; Fouille de données ; Apprentissage automatique.

Abstract

Knowing the uses of a digital library like Gallica requires to renew the survey methods traditionally used by the library, by semi-automatized exploration of server data. This exploration requires mathematical models that make the dialogue between researchers and library professionals more difficult. A research project conducted by the National Library of France and Télécom ParisTech on the Gallica logs shows the possibility of inscribing the data mining in a dialogue where researchers and professionals try to enlighten each other on the decisions to be made to conduct a relevant analysis. It also highlights the importance of combining scientific methods to understand online uses, none of which can claim to be self-sufficient.

Keywords

Digital Library; Use; Web; Heritage; Data Mining; Machine Learning.

Resumen

Conocer los usos de una biblioteca digital como Gallica requiere renovar los dispositivos de encuesta tradicionalmente utilizados por la biblioteca, mediante la exploración semiautomatizada de los datos del servidor. Esta exploración utiliza modelos matemáticos que dificultan el diálogo entre investigadores y profesionales de la biblioteca. Un proyecto de investigación llevado a cabo por la Biblioteca Nacional de Francia y Télécom ParisTech en los registros de Gallica muestra la posibilidad de inscribir la minería de datos en un diálogo en el que investigadores y profesionales intentan informarse mutuamente sobre las decisiones que deben tomarse para conducir un análisis relevante. También destaca la importancia de combinar métodos científicos para comprender los usos en línea, ninguno de los cuales puede afirmar ser autosuficiente.

Palabras clave

Biblioteca Digital; Uso; Web; Patrimonio; Minería de Datos; Aprendizaje Automático.

Introduction

Face au rythme accéléré des innovations sociotechniques et à l'éclatement des pratiques numériques, construire une vision globale des usages d'une bibliothèque numérique comme Gallica (gallica.bnf.fr) constitue un véritable défi qui nécessite de mobiliser de nouvelles méthodes, en particulier celles permettant d'analyser des données de masse. Ces méthodes requièrent des compétences nouvelles par rapport à celles mobilisées traditionnellement dans l'univers de la connaissance des pratiques culturelles en général et des bibliothèques en particulier. C'est là une mutation importante. Le domaine des « études de publics », qui s'est développé dans des établissements tels les bibliothèques et musées depuis les années 1980, au confluent de plusieurs disciplines (sociologie, ethnologie, marketing, etc.), était traditionnellement le lieu d'un dialogue relativement équilibré entre les professionnels de ces établissements et les experts en charge de mener l'enquête, qu'ils soient chercheurs académiques ou consultants. Si des difficultés se présentaient, elles relevaient rarement

de la compétence scientifique des interlocuteurs, les méthodes mobilisées, telles que l'enquête par questionnaire, l'observation ou l'entretien, étant facilement appropriables par tous. Cela ne signifiait bien entendu pas que tous pouvaient mener l'enquête, mais que tous pouvaient participer à la rédaction d'un questionnaire, comprendre l'enjeu des questions posées aux publics et se représenter le type de traitements réalisés sur les données collectées. Une telle participation au processus d'enquête est rendue plus difficile dans le cadre de la fouille de données (*data mining*), indispensable désormais pour la connaissance des usages en ligne. Un projet de recherche inédit sur les logs de connexion à Gallica, conduit dans le cadre d'un partenariat entre la Bibliothèque nationale de France (BnF) et Télécom ParisTech, témoigne pourtant de la possibilité d'inscrire la fouille de données dans un dialogue où chercheurs et professionnels s'efforcent de s'éclairer mutuellement sur les décisions à prendre pour conduire une analyse pertinente. Nous poserons dans un premier temps le cadre institutionnel et méthodologique dans lequel s'inscrit la connaissance des usages de Gallica, avant de présenter les apports de l'apprentissage automatique à cette connaissance et la manière dont ils peuvent être intégrés à la réflexion qu'une bibliothèque comme la BnF porte sur son offre en ligne.

Un contexte institutionnel singulier

La BnF a une tradition ancienne d'enquêtes auprès des usagers de sa bibliothèque numérique Gallica, l'une des plus grandes librement accessibles sur le web, avec plus de quatre millions de documents patrimoniaux numérisés et près de quinze millions de visites par an. Il n'y a là rien d'original par rapport à tout établissement culturel qui développe des services en direction d'un public qu'il doit s'efforcer de connaître. Depuis les années 1980, « connaître ses publics » est devenu, comme le rappelle Olivier Donnat, une « *figure rhétorique obligée pour la plupart des responsables culturels* » (Donnat, 2016, p.6). Mais si l'objet est en apparence simple, la manière dont il sera connu l'est beaucoup moins. Se pose la question non seulement des méthodes convoquées, mais aussi et surtout des compétences mobilisées et de leur place dans l'organisation : qui a l'initiative de commander une telle étude, à quelle fin et avec quelles formes d'implication dans sa réalisation ? Le champ d'activité que l'on nomme sans beaucoup de précision « étude de publics », d'une importance pourtant stratégique pour les établissements culturels, est rarement interrogé en ces termes, ceux de ses conditions de réalisation. Dans le cas des bibliothèques, le précieux guide *Mener l'enquête. Guide des études en bibliothèque* (Evans, 2011) laisse par exemple ouverte la question des acteurs des études, ceux-ci variant considérablement d'une institution à l'autre, ne serait-ce que pour des questions de moyens. Ce guide s'adresse en effet à « des non spécialistes qui travaillent dans des bibliothèques [...]. Que ces professionnels soient amenés à réaliser eux-mêmes un projet d'étude, en se transformant en enquêteurs pour l'occasion, ou que leur rôle se limite à accompagner ce projet en en confiant la réalisation à un tiers extérieur (prestataire spécialisé, stagiaire ou autre) » (Evans, 2011, p.9). Deux cas de figure sont donc envisagés pour la réalisation des études : dans le premier cas, le bibliothécaire peut « à l'occasion » se faire lui-même enquêteur, en s'appropriant sans trop de difficulté quelques principes et outils élémentaires de la sociologie ; dans le second cas, la compétence est déléguée à un tiers – ce qui recouvre en fait deux situations très différentes selon qu'il existe ou non, au sein de l'organisation concernée, une instance en charge du pilotage des « études de publics ». Si celle-ci existe, il convient alors de réfléchir à ses compétences, ses missions et sa place dans l'organigramme : des éléments structurels qui ont des conséquences sur la manière de conduire des études.

Dans le cas de la BnF, les « études de publics » sont validées dans leur principe et accompagnées tout au long de leur réalisation – de leur instruction à la diffusion des résultats –, par une instance en charge de définir chaque année un programme d'études prospectives et d'évaluation sur les activités

de la Bibliothèque : la délégation à la Stratégie et à la recherche, directement rattachée à la direction générale. Ce programme d'étude, validé en comité de direction, est instruit à partir des besoins exprimés par l'ensemble des services et des propositions de ladite délégation. Dans cette situation propre à la BnF, l'instance en charge de la connaissance des publics est donc distincte des instances opérationnelles qui ont besoin de cette connaissance pour faire évoluer leurs services, fidéliser ou conquérir de nouveaux publics. Une telle distinction ne se retrouve pas toujours dans les organisations qui préfèrent désormais lier la démarche de connaissance des publics à celle de leur conquête. Citons, à titre d'exemple, la Réunion des musées nationaux–Grand Palais (RMN-GP) qui a créé une « cellule études et marketing transversale » (Babault, Lévy-Fayolle, 2016, p.61) réunissant les deux fonctions. Cette fusion organisationnelle peut avoir deux conséquences sur la connaissance des publics : soit celle-ci se limite aux publics les plus influents et les plus engagés, sur lesquels le marketing de l'offre cherche alors à s'appuyer ; soit elle est d'abord l'occasion de créer du lien avec les publics, collecter des données personnelles, communiquer sur de nouvelles offres, etc. Les études de publics connaissent ainsi depuis plusieurs décennies une double évolution qui se traduit tout d'abord par le passage d'une logique de la représentativité (connaître de la manière la plus rigoureuse possible qui sont les publics d'un service ou d'un produit) à une logique de l'influence (qui sont, au sein des publics, ceux qui ont le plus d'influence) ; et le passage ensuite de la connaissance de ses publics à la volonté de les faire participer à l'élaboration de l'offre (Beaudouin, Denis, 2014, p.27).

Cette double évolution des études peut cependant réduire le temps consacré à l'activité de connaissance en elle-même, privilégiant le résultat rapide sur la fiabilité du chemin qui y conduit, et limitant le champ d'exploration des usages à quelques cibles ou signaux forts. Pour éviter ces écueils, la BnF a choisi depuis 2013 d'inscrire son activité de connaissance des publics en ligne dans le temps long d'un partenariat de recherche avec Télécom ParisTech, grande école spécialisée dans les technologies de l'information et des télécommunications. Ce temps long est nécessaire à l'apprivoisement mutuel de cultures professionnelles différentes et à l'élaboration de problématiques communes où les questions des chercheurs et les besoins d'un établissement comme la BnF espèrent trouver un terrain d'entente, sans décision déjà induite, dans un esprit de concertation et de respect mutuel. Cette rencontre n'est jamais acquise – le présent projet sur l'analyse des logs en témoigne – mais prend toujours la forme d'une négociation, portée par une conviction : le respect de l'autonomie de la recherche n'est pas un obstacle mais la condition de production de résultats fiables et utiles. Ce partenariat a donné naissance au Bibli-Lab, le « Laboratoire d'étude des usages du patrimoine numérique des bibliothèques », qui abrite des projets élaborés, financés et pilotés conjointement par la BnF et Télécom ParisTech, pour une durée qui varie entre six mois et trois ans, laissant une place importante à l'expérimentation.

La connaissance des usages de Gallica : un problème de méthode

Comme la plupart des études internationales sur les bibliothèques numériques, les études sur Gallica se sont inscrites traditionnellement dans le paradigme de la sociologie des « usages », au sens où il s'agit de mesurer un écart entre un usage prescrit par un dispositif et son déplacement par l'appropriation d'un « usager » considéré comme un acteur relativement autonome (Jahjah, 2017 ; George, 2012).

Créée fin 1997 avec 20 000 documents accessibles, Gallica a fait l'objet dès 2003 d'un premier projet de recherche mené en partenariat avec France Télécom R&D. Croisant les méthodes de l'enquête en ligne, de l'analyse de trafic sur panel et de l'entretien qualitatif, ce projet mettait en avant une approche « *centrée utilisateur complète, qui reste rarement mise en œuvre dans les études d'usage d'envergure sur le Web* » (Assadi et al., 2003, p.2). Force est cependant de constater que ce sont les méthodes traditionnelles d'enquêtes qui ont dominé par la suite l'approche des usages de Gallica,

avec cinq études aux objets divers conduites entre 2007 et 2011 : entretiens semi-directifs, individuels ou collectifs, et questionnaires en ligne. La tendance était alors, consciemment ou non, d'étudier les usagers de la bibliothèque numérique comme ceux de la bibliothèque physique : on entrerait dans Gallica comme on pousse la porte du site François-Mitterrand ou Richelieu. S'y retrouveraient par conséquent les mêmes finalités (lire, se documenter) et les mêmes chaines d'actions (chercher un document dans le catalogue, le consulter), dans une temporalité à peine resserrée (45 % des répondants à une enquête en ligne en 2011 déclaraient passer plus d'une demi-heure sur Gallica (GMV, 2011, p.21)). Cette tendance à recourir aux méthodes traditionnelles était alors largement partagée si l'on en croit la recension réalisée par Carol Tenopir de quelques deux cent études sur les usages des bibliothèques numériques dans le monde anglo-saxon (Tenopir, 2003).

Si elles demeurent des sources d'information irremplaçables, ces méthodes traditionnelles se heurtent cependant au caractère à la fois majoritairement furtif (proche du « braconnage » : je regarde, je prends, je m'en vais) et entrelacé (multi-support, multi-activité) des usages d'une interface comme celle de Gallica, comme le vérifient à la fois l'analyse de l'audience de l'interface et l'observation de l'activité réelle (Rollet et al., 2017). Quelques chiffres soulignent la nécessité de changer de référentiel par rapport aux usages d'une bibliothèque physique : 50 % des visites de Gallica font moins de 12 secondes ; 30 % ne font qu'une seule requête et seulement 8 % des sessions consultent plus de quatre documents uniques. Ces difficultés, bien connues de la sociologie du web (Beuscart et al, 2016), limitent la portée des informations collectées via les méthodes d'enquête traditionnelles – qualitatives (entretiens) ou quantitative (questionnaires en ligne). Les usagers ont de plus en plus de mal à raconter, en entretien, le détail de leurs pratiques. Et leur propension à répondre à des enquêtes en ligne décroît à mesure qu'augmentent la rapidité et l'habitude avec lesquelles ils font des « choses » sur le web.

Ces lacunes des enquêtes traditionnelles – qui ne les invalident pas pour autant – avaient été relevées au lancement du Bibli-Lab par Valérie Beaudouin et Jérôme Denis, dans un travail sur les enjeux théoriques et méthodologiques de la connaissance des usages de Gallica (Beaudouin et Denis, 2014). Ces chercheurs préconisaient non pas une substitution mais une « articulation des outils et des formats d'investigation » (ibid., p.26). Cette articulation touchait en fait de près à l'organisation même de la production d'information sur les publics en ligne au sein de la BnF. Ces publics sont en effet rendus présents à travers des formes d'engagement variées, qui se trouvent gérées ou analysées par différences services de la BnF : 1) connaissance sociologique des publics par des chargés d'étude, 2) suivi de l'audience par des informaticiens, 3) veille sur les réseaux sociaux par des *community managers*, 4) relation-client par des responsables-produits. La conclusion était que la BnF n'avait pas tant un déficit d'informations sur ses publics en ligne – elle en avait même beaucoup plus que ce que chaque service isolé imaginait –, qu'un déficit d'articulation raisonnée entre ces informations. Rapprocher la connaissance sociologique des publics avec le suivi d'audience invitait en particulier à prendre en compte, aux côtés de la parole d'un nombre forcément limité d'usagers, l'ensemble des connexions à Gallica (15 millions de visites par an). Cela voulait dire par la même occasion décentrer l'analyse et les notions traditionnellement manipulées : passer de la notion anthropologiquement et sociologiquement connotée d'» usager » à celle d'» usage ». Autrement dit, identifier la cohérence interne des usages, sans présumer de celle de l'usager supposé rationnel et régulier dans ses engagements – un même usager pouvant avoir plusieurs usages, à des moments différents et pour des motivations différentes (Beaudouin et Denis, 2014, p.27).

La voie était ainsi ouverte pour une analyse inédite des logs de connexion à Gallica, en leur appliquant des méthodes de fouille de données, incluant de l'apprentissage automatique (*machine learning*). En effet, si le recours aux logs pour améliorer l'expérience-usager des bibliothèques numériques n'est pas une idée nouvelle, les études internationales que nous avons pu consulter en amont se limitent à des traitements statistiques simples, descriptifs, sans visée prédictive (cf.

Ceccarelli et al., 2011, et les travaux qu'ils citent dans ce domaine). La perspective neuve était ici de prendre l'ensemble des connexions à un service pour y repérer des similitudes dans les enchaînements d'événements : quand on fait *a*, quelle est la probabilité de faire *b* ?

Des informations inédites, complémentaires des autres approches

Les logs de connexion à un site web sont des fichiers qui contiennent toutes les requêtes reçues par les serveurs hébergeant le site. Dans le cas de Gallica, ils n'étaient initialement conservés par la BnF qu'à des fins de sécurité et d'évaluation de la qualité de service. Un certain nombre d'ajustements techniques ont donc dû être opérés pour la présente recherche : conservation de l'agent-utilisateur (*user-agent*) et du site afférent (*referer*) dans les logs, résolution d'un problème d'incomplétude des logs, anonymisation des données pour des raisons juridiques et éthiques. Après ces ajustements acceptés et réalisés par le département des systèmes d'information de la BnF – impliqué dès le début du projet –, les chercheurs de Télécom ParisTech ont pu disposer de fichiers contenant des informations importantes pour la connaissance des usages de Gallica : l'adresse I.P. (qui fait office d'identifiant unique d'une connexion, anonymisé pour le présent projet), la date et l'heure (à la seconde près) de la requête, la provenance de l'utilisateur (site référent), ou encore la requête http qui, dans le cas de l'appel d'un document de Gallica, contient son identifiant pérenne ARK. L'objectif, en traitant ces données de masse, n'était pas de connaître les usagers et leurs profils, accessibles seulement à travers des enquêtes déclaratives, mais d'identifier des types de navigation. De cette manière, la proposition de Valérie Beaudouin et Jérôme Denis (2014) de décentrer l'analyse de l'usager à l'usage, était bien respectée.

Durant quinze mois (avril 2016-juillet 2017), un chercheur en contrat postdoctoral, Adrien Nouvellet, encadré par quatre enseignants-chercheurs de Télécom ParisTech (Valérie Beaudouin, Florence d'Alché-Buc, Christophe Prieur et François Roueff) a eu en charge les missions suivantes : 1) nettoyer les logs de connexion (filtrage des robots internet) ; 2) structurer les logs de connexion en leur appliquant une notion de session (succession de requêtes d'une même adresse I.P. dont l'écart temporel n'excède pas 60 minutes) ; 3) définir les quatre actions-types utiles à l'analyse des sessions obtenues (consultation de la page d'accueil ; utilisation du moteur de recherche interne ; consultation d'un document dans l'interface ; téléchargement) ; 4) analyser les parcours d'usage en mettant au point un algorithme de partitionnement de données (ou *clusterisation*) qui permette de regrouper des sessions présentant des similitudes dans l'enchaînement et la durée des actions ; 5) analyser les types de documents consultés dans les sessions et les types d'action qui leur sont liés.

À un niveau général, cette analyse des logs a révélé une diversité de parcours que les entretiens et enquêtes en ligne – qui ne captent que les usagers les plus engagés –, ont tendance à fortement réduire : poids des sessions très courtes dans l'audience globale, forte fluctuation temporelle des thèmes les plus consultés, variation du nombre d'actions effectuées dans Gallica en fonction de la provenance web de l'internaute, etc. Ce niveau simplement descriptif se révèle cependant assez vite pauvre en informations, similaires à ce que fournissent les requêtes pré-codées des outils de mesure d'audience. Aller au-delà des indicateurs de trafic, c'est être en mesure de passer d'un objectif général de connaissance, à la formulation souvent vague (mieux connaître les usages de l'offre, savoir ce qu'il conviendrait d'améliorer) et à la formulation de véritables hypothèses de recherche, qui touchent à la finalité même de l'offre et à ce que l'institution met en œuvre pour l'atteindre : quelles postures et actions sont visées par les développements réalisés ? Ce moment réflexif pour l'institution est forcément critique, tant l'usage présumé d'une offre ou d'un produit peut être incertain, contradictoire, résultat de compromis et de logiques hétérogènes. Dans cette difficulté à formuler ce que l'on veut savoir précisément, l'objet technique peut donner l'impression de s'être développé tout seul, sans « script » explicitement formulé. La difficulté du dialogue avec les chercheurs est ici

redoublée par l'opacité pour un observateur extérieur des modèles mathématiques utilisés pour explorer les données de masse : le non expert a des difficultés à se représenter le type de résultats susceptibles d'être produits tant qu'il n'en a rien « vu » et peine donc à formuler une demande dont l'expert a malgré tout besoin pour avancer, organiser ses données et montrer « quelque chose » qui relance le dialogue.

« Mais que voulez-vous savoir, précisément ? » fut la question la plus fréquemment posée au démarrage de ce projet de recherche, poussant la BnF à formuler deux hypothèses de recherche, qu'il convenait de mettre à l'épreuve des faits :

1) Hypothèse de la diversité documentaire : une bibliothèque numérique comme Gallica favorise l'exploration des fonds numérisés dans toute leur diversité documentaire (presse, livres, manuscrits, estampes, photographies, etc.) et leur profondeur historique, contrairement à ce qui est constaté dans les salles de lecture où domine la consultation de documents très récents, dans une logique monodisciplinaire (Pardé, 2015) ;

2) Hypothèse de l'impact de la médiation : les actions de médiation (création de pages de médiation présentant des collections particulières, éditiorialisation de la page d'accueil et usage des réseaux sociaux) favorisent la découverte et l'exploration des fonds de Gallica.

Ces deux hypothèse ont conduit à enrichir les logs par les métadonnées descriptives des documents présentes dans l'entrepôt OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting : protocole de moissonnage des données). Ces métadonnées incluent, entre autres, la date d'édition et le type de document. Il a également été décidé de distinguer une action-type « consultation d'une page de médiation » (présentation des collections et blogue), au sein des actions identifiées comme pertinentes pour l'analyse – pour mémoire, les quatre initialement définies étaient : consultation de la page d'accueil ; utilisation du moteur de recherche interne ; consultation d'un document dans l'interface de Gallica ; téléchargement.

Ces deux hypothèses sont loin d'avoir été confirmées par les analyses, qui ont apporté à l'institution des informations précises et inédites. Nous nous contentons ici de résumer les principaux résultats, l'intégralité du rapport étant librement accessible (Nouvellet et al., 2017).

Hypothèse de la diversité documentaire : une faible diversité des types de documents consultés au sein d'une session

Ce résultat est une surprise : malgré les facilités d'exploration qu'offrent les interfaces du web et la sérendipité si souvent mise en avant, les consultations de Gallica restent largement monotypes. C'est le cas de 45 % des sessions à plus de cinq documents, avec une prédominance des sessions ne consultant que des fascicules de presse ou des monographies. Ces sessions à plus de cinq documents, pourtant plus longues que la moyenne, reproduisent une logique de consultation en « silos », à l'image de l'organisation des collections et des pratiques de recherche encore cloisonnées, comme l'avait vérifié l'étude en 2012 des demandes de documents en Rez-de-jardin (Pardé, 2015). Un défi pour l'interface de Gallica sera de favoriser une logique de rebond d'un type à l'autre (par exemple : d'un manuscrit d'Apollinaire à l'écoute de sa voix). Seules 3 % des sessions à plus de 5 documents explorent presque l'ensemble des types de documents.

Hypothèse de l'impact de la médiation : une médiation statique peu efficace

Si la conception d'un site web induit toujours une présomption d'usage « normal » (par exemple : page d'accueil > moteur interne > consultation de document), les clusters vérifient la très grande diversité des logiques de parcours dans Gallica. Dans le premier modèle de clusters obtenus, ne prenant en compte que la succession des actions, 53 % des sessions correspondent à des séquences de pure consultation de documents qui ne passent pas par la page d'accueil, ne téléchargent pas et

n'utilisent pas le moteur de recherche. Il convient donc pour Gallica de ne pas concevoir la page d'accueil comme la porte d'entrée principale, mais de faire au contraire de toute page, une porte d'entrée dans le site avec des propositions de parcours. En l'état actuel, les pages de présentation des collections apparaissent dans un cluster unique, de faible amplitude (2,5 %), vérifiant que ces pages ne sont pas sur la route de la plupart des *gallicanautes* ; leur consultation obéit à un comportement distinct de tous les autres observés. En revanche, l'impact sur l'audience de Gallica des actions de médiation sur les réseaux sociaux est avéré et Facebook est bien représenté dans les sites référents. Une étude sur le type de lien vers Gallica présent dans les publications a d'ailleurs montré que celui-ci avait des conséquences sur le nombre de « clics » : ainsi, un lien actif dans l'image engendre 25 fois plus de visites sur Gallica qu'un lien actif dans le texte (avec indication de l'URL). Ce résultat a incité l'équipe en charge de la page Facebook à modifier ses modalités de publication.

Des calculs en contexte de grande incertitude

La durée importante de cette recherche (15 mois) s'explique en partie par le temps nécessaire à tous les acteurs impliqués pour parvenir à se comprendre et assimiler le type de calcul susceptible d'être fait avec les données afin d'orienter la recherche. En effet, les opérations propres à la fouille de données, tel l'apprentissage automatique, ne se racontent pas facilement, contrairement à celles de l'ethnographie ou de la sociologie. En ce sens, le « big data » rend encore plus aigu le problème traditionnel du rapport entre recherche et institution, expertise et usage de cette expertise : il ne faudrait pas seulement « ouvrir les données », selon un leitmotiv contemporain qui se veut démocratique, il faudrait « ouvrir les modèles », au moins comprendre ce à quoi ils sont sensibles. Par exemple : l'ajout, dans une deuxième partie du travail, du facteur temps dans les clusters de sessions a été décisif pour l'interprétation finale, permettant de rapprocher des sessions au préalable éclatées mais qui présentent la même « silhouette » temporelle.

Sans doute, les difficultés éprouvées dans ce dialogue entre *data scientists* et professionnels des bibliothèques ne sont-elles que la résurgence d'un problème qui se pose depuis longtemps autour des sciences de la nature – où l'expert et le citoyen, le savant et le politique ont vu leurs positions respectives bouleversées par la spécialisation des savoirs. Mais ce problème se posait traditionnellement sur le plan de l'éthique, ou encore du bon usage de la science (Moreau, 2008), alors qu'il se pose ici résolument au plan épistémologique, qui est celui de la manière dont la science procède, construit et valide des hypothèses. La grande richesse statistique des « *big data* » est en effet extrêmement sensible aux artefacts techniques et pauvres en explications : ce paradoxe est inhérent aux nouvelles méthodes sociologiques basées sur les traces (Beuscart et al, 2016). Ce contexte de grande incertitude sur l'activité réellement décrite par les données est sans doute une clé du problème : la manipulation des données de masse a besoin de recourir à des informations tierces, détenues par d'autres acteurs, et le *data scientist* ne saurait travailler seul.

Conscient de ce besoin d'informations supplémentaires, le présent projet a reposé sur un dispositif original de gouvernance où compétences et méthodes ont systématiquement été croisées. Aux compétences initiales dans le domaine de la fouille de donnée (maîtrise des méthodes permettant d'explorer des données de masse et d'en extraire des informations susceptibles de décisions) se sont ajoutées des compétences de trois types : 1) informatiques (connaissance du format des données d'origine et de leurs lacunes éventuelles : manière dont elles ont été collectées et conservées) ; 2) bibliothéconomiques (connaissance des objets concernés par les traces d'usage : la « collection numérique » et sa place aujourd'hui dans les pratiques savantes) ; 3) celles enfin issues d'autres disciplines scientifiques analysant l'activité humaine, en partant du principe que les nouvelles approches ne se substituent pas aux plus traditionnelles (Beaudouin et Denis, 2014), et que les

représentants des sciences dites de la « complexité » et ceux des sciences sociales doivent collaborer (Lazega et Prieur, 2014).

L'analyse quantitative a ainsi été enrichie dès le départ de données issues d'autres enquêtes et observations qualitatives, en particulier une série d'entretiens exploratoires (Beaudouin et al., 2016) et une vidéo-ethnographie de *gallicanautes* (Rollet et al., 2017). Cet enrichissement s'est révélé nécessaire tout d'abord en amont, pour construire les premières définitions nécessaires à la modélisation : les notions de « session » et d'» action ». La vidéo-ethnographie avait en effet vérifié l'existence de très longues consultations d'une simple vue, ce qui a conduit à revoir la définition d'une session par rapport à l'état de l'art : si celui-ci considère qu'une session sur un site web se termine lorsque le temps entre deux requêtes excède 30 minutes, il a été décidé de porter ce temps pour Gallica à 60 minutes. Autre exemple notable d'articulation entre les approches qualitatives et quantitatives : l'usage du moteur de recherche de Google pour chercher dans Gallica, y compris à l'intérieur d'une même session, avait été repéré au préalable dans les entretiens exploratoires (Beaudouin et al., 2016, p.20) ; il a donc été décidé, en cours de recherche, d'ajouter aux cinq actions caractéristiques de l'usage de Gallica (cf. *supra*) une sixième : « Je fais une recherche via Google » – ce qui souligne au passage le caractère nécessairement évolutif des définitions, et donc des modèles. En aval, ce croisement des méthodes et des informations a été tout aussi décisif pour donner du sens aux résultats et ne pas attribuer à des acteurs ce qui ne relève que du dispositif technique.

La condition pour qu'un tel dialogue entre compétences et méthodes soit fructueux est l'existence d'une compréhension partagée des analyses conduites. Si les arcanes des modèles n'ont pas vocation à être compris en profondeur par toutes les parties prenantes, il est indispensable de prévoir des dispositifs de visualisation des résultats et de retour aux données collectées, qui deviennent une ressource pour le travail collectif de validation et d'interprétation des résultats. Aux traitements statistiques et à leurs lignes de code, il faut être en mesure de faire correspondre des « images », l'imagination étant, comme le rappelle Kant, une faculté constitutive de la connaissance qui permet l'articulation du perçu et du conçu.

Conclusion

L'application aux logs de Gallica des méthodes de fouille de données permet de compléter les méthodes traditionnelles en prenant en compte la très forte hétérogénéité des parcours, incluant en particulier les usages furtifs, faiblement motivés ou engagés, difficilement accessibles par les questionnaires en ligne ou les entretiens. Au-delà de ce que fournissent les statistiques de consultation descriptives, les méthodes d'apprentissage automatique permettent d'identifier dans la masse des connexions des régularités dans les chaînes d'actions et d'isoler des sessions-types dont on peut mesurer le poids au sein de l'ensemble des sessions. Par rapport à l'approche classique orientée « usagers », une telle recherche fait l'économie, au moins dans un premier temps, de « présupposés de cohérence, de rationalité et de constance chez les [usagers] », évitant de trop vite projeter sur les usages l'unité de « personnes », aux caractéristiques sociologiques et aux motivations précises (Beaudouin et Denis, p. 29). Elle permet ainsi de poser des constats macroscopiques solides sur ces usages, qui viennent dans certains cas contredire ceux imaginés par ses concepteurs (par exemple : la diversité des recherches documentaires) et même ceux déclarés par les usagers eux-mêmes (durée de leurs sessions, manière dont ils rationalisent à posteriori leurs stratégies de recherche, etc.).

À l'issue de cette recherche, le souhait pour l'institution de pérenniser certains traitements, afin de pouvoir les rejouer ponctuellement – dans la perspective en particulier de mesurer l'évolution des usages suite à une évolution majeure de l'interface –, a été exprimé et demeure à l'étude, ce qui pose

aussitôt la question des compétences en science des données dont un établissement comme la BnF doit se doter. Le souhait de disposer d'outils fournissant des résultats simples et rapides pour la prise de décision ne doit cependant pas occulter l'importance d'inscrire la fouille de données dans une démarche de recherche où les éléments structurant chaque étape, en particulier les hypothèses que l'on formule et les définitions temporaires que l'on se donne, doivent être discutés et peuvent être à tout moment modifiés en fonction d'autres observations. L'importance de ces autres observations, afin de multiplier les points de vue sur le même objet et lui redonner le cas échéant son épaisseur sociale (Beuscart, 2017), est un appel à l'interdisciplinarité : la fouille de données n'a ici de sens que si elle est au service d'hypothèses qui sont construites ailleurs, en particulier via les enquêtes qualitatives auprès des usagers, conduites par les sociologues ou les ethnologues, mais aussi via les échanges avec les professionnels des bibliothèques qui sont aussi un point de contact avec les usagers et un lieu d'expertise sur les collections numériques. Il convient enfin de rappeler que certaines hypothèses ne peuvent être testées dans les modèles statistiques, comme par exemple la distinction entre des parcours de recherches ciblées et une exploration libre, très prégnante pour les usagers (Beaudouin et al., 2016 ; Auray, 2017). C'est donc bien une complémentarité des approches qui doit être défendue.

Références bibliographiques

Assadi Houssem, Beauvisage, Thomas ; Lupovici, Catherine ; Cloarec, Thierry (2003), « Users and Uses of Online Digital Libraries in France », p. 1-12, in Koch, Traugott ; Sølvberg Ingebord T. (dir.), Research and Advanced Technology for Digital Libraries, ECDL 2003, *Lecture Notes in Computer Science*, vol. 2769, Springer, Berlin, Heidelberg : Springer.

Auray, Nicolas (2017), *L'Alerte ou l'enquête : Une sociologie pragmatique du numérique*, coll. « Sciences sociales », Paris : Presses des Mines.

Babault, Gaëlle ; Lévy-Fayolle, Florence (2016), « Les usages en ligne autour des expositions du Grand Palais », *Culture et Recherche*, n°134, p. 58-61.

Beaudouin, Valérie et Denis, Jérôme (2014), « Observer et évaluer les usages de Gallica. Réflexion épistémologique et stratégique », Rapport de recherche, BnF, Telecom ParisTech, [en ligne], Consulté le 16 février 2018, https://halshs.archives-ouvertes.fr/halshs-01078530/document.

Beaudouin, Valérie ; Garron, Isabelle ; Rollet, Nicolas (2016), « Je pars d'un sujet, je rebondis sur un autre : pratiques et usages des publics de Gallica », étude qualitative exploratoire, Rapport final de la phase 1 du projet « Mettre en ligne le patrimoine : transformation des usages, évolutions des savoirs ?, Bibliothèque nationale de France, labex Obvil, Télécom ParisTech, [en ligne], Consulté le 19 février 2018, https://hal.archives-ouvertes.fr/hal-01709238/document.

Beuscart, Jean-Samuel ; Dagiral Eric ; Parasie, Sylvain (2016), *Sociologie d'internet*, coll. « Cursus », Malakoff : Armand Colin.

Beuscart, Jean-Samuel (2017) « Des données du Web pour faire de la sociologie... du Web ? », in Menger, Pierre-Michel ; Paye, Simon (coord.), *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus*, nouvelle édition, Paris : Collège de France, [en ligne], Consulté le 10 mai 2018, http://books.openedition.org/cdf/4987, DOI : 10.4000/books.cdf.4987.

Ceccarelli, Diego ; Gordea, Sergiu ; Lucchese, Claudio ; Nardini, Franco Maria ; Tolomei, Gabriele (2011), « Improving Europeana search experience using query logs », p. 384-395, in Gradmann, Stefan ; Borri, Francesca ; Meghini, Carlo ; Schuldt, Heiko (coord.), *Research and Advanced Technology for Digital Libraries, International Conference on Theory and Practice of Digital Libraries*, Berlin Heidelberg : Springer-Verlag.

Donnat, Olivier (2016), « La question du public, d'un siècle à l'autre », *Culture et Recherche*, n°134, p. 6-8.

Evans, Christophe (coord.) (2011), *Mener l'enquête. Guide des études de publics en bibliothèque*, coll. « La boîte à outils », Villeurbanne : Presses de l'Enssib.

George, Éric (2012), « L'étude des usages des TIC au prisme de la recherche critique en communication », p. 25-63, in Vidal, Geneviève (coord.), *La sociologie des usages. Continuités et transformations*, coll. « Environnement et services numériques d'information », Cachan : Lavoisier.

GMV (2011), *Évaluation de l'usage et de la satisfaction de la bibliothèque numérique Gallica et perspectives d'évolution*, Rapport détaillé, Bibliothèque nationale de France.

Jahjah, Marc (2017), « État de l'art théorique, méthodologique et critique sur les usages et les pratiques », Rapport, Phase 1 du projet « Mettre en ligne le patrimoine : transformation des usages, évolutions des savoirs ?, Bibliothèque nationale de France, labex Obvil, Télécom ParisTech, [en ligne], Consulté le 19 février 2018, http://www.enssib.fr/bibliotheque-numerique/documents/67532-etat-de-l-art-theorique-methodologique-et-critique-sur-les-usages-et-les-pratiques.pdf.

Lazega, Emmanuel ; Prieur, Christophe (2014), « Sociologie néostructurale, disciplines sociales et systèmes complexes », *Revue Sciences/Lettres*, t. 2, [en ligne], Consulté le 17 février 2018, http://rsl.revues.org/455, DOI : 10.4000/rsl.455.

Moreau, Didier (2008), « Le Savant et la pédagogue : despotisme ou démocratie ? », p.347-363, in Mustière, Philippe ; Fabre, Michel (coord.), *Jules Verne, le partage du savoir, Actes du colloque international*, École centrale, Nantes : Coiffard.

Nouvellet, Adrien ; Beaudouin, Valérie ; D'Alché-Buc, Florence ; Prieur, Christophe ; Roueff, François (2017), « Analyse des traces d'usage de Gallica : Une étude à partir des logs de connexions au site Gallica », Rapport de recherche, Télécom ParisTech, Bibliothèque nationale de France, [en ligne], Consulté le 28 février 2018, <https://hal.archives-ouvertes.fr/hal-01709264>.

Pardé, Thierry (2015), « Les usages documentaires dans une bibliothèque de Recherche », *Bulletin des bibliothèques de France* (BBF), n° 5, p. 112-119, [en ligne], Consulté le 17 février 2018 : http://bbf.enssib.fr/consulter/bbf-2015-05-0112-002.

Rollet, Nicolas ; Beaudouin, Valérie ; Garron, Isabelle (2017), « Vidéo-ethnographie des usages de Gallica », Rapport final de la phase 2 du projet « Mettre en ligne le patrimoine : transformation des usages, évolutions des savoirs ?, Bibliothèque nationale de France, labex Obvil, Télécom ParisTech, [en ligne], Consulté le 19 février 2018, https://hal.archives-ouvertes.fr/hal-01709210.

Tenopir, Carole (2003), « Use and Users of Electronic Library Resources An Overview and Analysis of Recent Research Studies », *Council on Library and Information Resources*, Washington DC, [en ligne], Consulté le 10 mai 2018, https://clir.org/wp-content/uploads/sites/6/pub120.pdf.

TMO régions (2017), « Enquête auprès des usagers de la bibliothèque numérique Gallica », Rapport d'enquête, Bibliothèque nationale de France, [en ligne], Consulté le 17 février 2018, http://www.bnf.fr/documents/mettre_en_ligne_patrimoine_enquete.pdf

L'éditorialisation des données aux bornes des API : Enjeux et perspectives pour une analyse empirique

The Editorialization of the data at APIs' bounds: Issues and perspectives for an empirical analytic

Editorialización de datos en los terminales de la API: Temas y perspectivas para el análisis empírico

Article inédit, mis en ligne le 15 novembre 2018.

Cette réflexion est associée à la conduite de projets menés dans le cadre du WP4 du Grenoble Alpes Data Institute, supporté par l'Agence Nationale pour la Recherche - Investissements d'avenir (ANR-15-IDEX-02).

Jean-Marc Francony

Jean-Marc Francony est Maître de conférences en sciences de l'information-communication à l'Université Grenoble Alpes. Il est membre permanent du laboratoire PACTE en sciences sociales. Ses recherches portent sur les dispositifs info-communicationnels numériques, l'analyse des pratiques et des usages sociaux de l'Internet, ainsi que sur le développement de méthodes digitales dans un contexte de données massives.

Plan de l'article

Résumé

Les interfaces publiques d'accès aux données contribuent à la définition d'un écosystème de services pour les plateformes du Web 2.0. La publicisation de ces données impose pour ces acteurs de l'économie numérique de concevoir une ouverture graduelle monétisable et d'opérer une réduction informationnelle qui s'apparente à un processus d'éditorialisation. La compréhension de ces mécanismes est essentielle dans la conduite de projets d'analyse de données et de recherches empiriques ou pour nourrir les questionnements critiques, méthodologiques et épistémologiques en SIC.

Mots clés

API, Twitter, *Internet Studies*.

Abstract

Applicative public interface to data are used by Web 2.0's platforms to create an ecosystem of services. For these actors of the digital economy, Data publication involve to conceive a monetisation and a gradual opening and to carry out informational reduction which is similar to an editorial process. The understanding of these mechanisms is essential in the conduct of data analysis and empirical research or to feed critical, methodological and epistemologicial questions in ICS.

Keywords

API, Twitter, Internet Studies.

Resumen

Las interfaces públicas de acceso a datos contribuyen a definir un ecosistema de servicios para las plataformas Web 2.0. La publicación de estos datos requiere que estos actores de la economía digital diseñen una apertura gradual y monetizable y realicen una reducción informativa que se asemeje a un proceso de editorialización. La comprensión de estos mecanismos es esencial para llevar a cabo proyectos analíticos e investigaciones empíricas o para alimentar cuestiones críticas, metodológicas y epistemológicas en el SIC.

Palabras clave

API, Twitter, Estudios en Internet

Introduction

Comme d'autres disciplines des sciences humaines et sociales, les sciences de l'information et de la communication (SIC) sont confrontées aux enjeux des flux de données massives du Web, et tout particulièrement ceux associés aux réseaux sociaux. Si la prise en compte de ces flux s'impose en tant qu'objet et ne fait pas débat, il en va tout autrement de leur exploitation et de leur interprétation (Bigot et al., 2016), (Paquienséguy et al., 2017). Les questions soulevées renvoient à des considérations méthodologiques mais aussi épistémologiques.

La genèse de l'expression Web 2.0 et les difficultés à la justifier rétrospectivement fonde une critique légitime de son usage scientifique (Bouquillion et Matthews, 2010). De plus, le terme Web 2.0 a étendu la définition du Web à l'ensemble des plateformes dites réseaux sociaux numériques qui ont émergé depuis 2000. S'il ne peut être question d'une notion caractérisant une pratique sociale ou un

modèle économique, ce jalon historique reste, à nos yeux, pertinent dans l'évolution des techniques de l'Internet. En effet, du point de vue de l'architecture informatique, les années suivantes voient un modèle de plateforme de services s'imposer comme standard du Web[1].

Dans le même temps, les flux de données rendus disponibles par les opérateurs du Web acquièrent de la valeur. Comprendre les modalités de leur fabrication et de leur publicisation est indispensable en amont de leur exploitation. La rétro-ingénierie informationnelle apporte un éclairage sur ces modalités. En suivant cette approche, nous nous intéressons aux flux de données publicisés dans les API de Twitter. L'objectif est à la fois de comprendre le processus de fabrication et ainsi la nature des données disponibles et leur portée, mais également de formuler des hypothèses sur les logiques de leur publication.

Les enjeux des API

Tim O'Reilly (O'Reilly, 2005) a eu le mérite d'énoncer les caractéristiques fonctionnelles et informationnelles désormais dominantes de ces plateformes de services. Les préconisations associées à cette conception encouragent les formes ouvertes et collaboratives de développements informatiques. La mise en œuvre d'interfaces de programmation applicatives ou API (*Applications Programming Interface*) est une méthode d'ouverture par les données visant le développement de services annexes et ainsi la croissance d'un écosystème périphérique à la plateforme[2]. L'avènement d'un Internet des Objets (*IoT,* pour *Internet of Things*, aussi identifié comme Web 4.0) conforte ces choix structurels et impose le recours aux API dans le développement de systèmes numériques complexes pour en garantir l'interopérabilité (Institut Montaigne, 2015).

Bien que l'information mise à disposition aux bornes des API se doive d'être le reflet de leur activité, il ne s'agit cependant pas, pour ces entreprises que sont les plateformes de services, de rendre accessibles les éléments d'un avantage concurrentiel sans contrepartie. Ainsi voit-on de plus en plus le développement des API publiques se décliner suivant différents modèles d'accessibilité plus ou moins étendue en fonction d'une échelle de rémunération associée. L'accès gratuit devient un mode standard dont les restrictions qualitatives et quantitatives sur le flux sont justifiées par l'existence de modes premiums payants offrant graduellement un accès plus performant.

Annoncées pour le printemps 2018, les nouvelles perspectives réglementaires européennes sur la protection des données personnelles (GDPR) incitent les acteurs du Web 2.0 à réduire toujours plus l'ouverture sur les données qu'ils élaborent, tout au moins à en maîtriser davantage leur diffusion. En effet, pour ces plateformes, il est indispensable d'intégrer les interactions possibles entre les données publicisées via les API avec celles issues des nouvelles possibilités de valorisation découlant du principe de portabilité des données personnelles (article 20-2).

Dans ce contexte actualisé, nous formulons l'hypothèse que le processus de réduction informationnelle opérée aux bornes des API publiques s'apparente à un processus d'éditorialisation qui en fixe la valeur d'usage et oriente le développement de l'écosystème de services.

L'accès aux données d'Internet relatives aux activités humaines constitue un enjeu de premier plan dans la compréhension de l'évolution sociétale ou des pratiques sociales des dispositifs numériques. L'usage intensif d'objets connectés au Web, la complexité croissante des interactions médiatisées

. . . . . .

1 Les travaux sur l'architecture orientée service (SOA) datent du début des années 2000.

[2] On évoque pour Twitter un trafic 10 fois supérieur via ses API que via sa plateforme Web. (http://avc.com/2007/09/biz-stone-on-re/)

rendent nécessaire l'accès aux données publicisées par les plateformes de services. De nombreux acteurs publics et privés s'emploient ainsi à utiliser les API publiques pour collecter et analyser les données. Cependant, la généralisation de la monétisation des flux crée des paliers entre l'accès gratuit et les modalités payantes dont les conséquences qualitatives sur les analyses méritent d'être soulignées.

L'exemple de Twitter est de ce point de vue emblématique. Cette plateforme du Web social est probablement l'une des plus étudiées. L'instantanéité de ses publications fait de Twitter un outil privilégié d'alerte lors de la survenue d'événements, ou de signalement lorsque de nouvelles publications sont en ligne. Cette caractéristique se double de l'avantage qu'offrent les API de Twitter publiques d'accéder en masse aux représentations numériques très structurées de tweets qui concentrent pour chacune d'elles une information qu'il serait fastidieux de collecter depuis la page Web de comptes Twitter. L'accessibilité et la diversité des données ainsi structurées éveillent l'intérêt d'analystes souhaitant produire des services ou de la connaissance à partir des flux de Tweets ou de leurs réseaux de diffusion.

En partant de l'exemple de Twitter, dont les quelque 500 millions de messages instantanés journaliers suscitent une convoitise, nous proposons de formaliser les restrictions informationnelles issues de l'éditorialisation des données, et nous envisageons les conséquences de ce processus du point de vue général de l'économie des données mais également du point de vue méthodologique et épistémologique des sciences de l'information et de la communication.

L'analyse des flux de données : l'exemple de Twitter

Au travers des conditions générales d'usages (CGU), de la définition fonctionnelle des API ou enfin de la construction du flux informationnel, le processus éditorial définit un cadre normatif pour l'analyse. Pour la recherche, la connaissance de ce cadre est une nécessité méthodologique. Elle permet d'évaluer l'étendue, la pertinence et la portée heuristique des collections envisagées. Dans le cas de Twitter, c'est la restriction en volume du flux délivré qui est généralement évoquée comme limitation de l'API gratuite. Mais la conséquence de cette limitation, établie à 1% du flux total, n'en reste pas moins floue pour la majorité des travaux de recherche (Morstatter et al., 2014). En novembre 2017, Twitter a récemment réaménagé son modèle économique et a profondément remanié l'accès à ses interfaces de données. L'offre se trouve enrichie d'accès historiques, géolocalisés ou encore d'informations volumétriques, selon des opérateurs premiums dont les performances, calibrées selon une grille tarifaire mensuelle[3], sont susceptibles de répondre aux besoins qualitatifs et quantitatifs d'entreprises émergentes.

Les modalités d'accès

Collecter des informations aux bornes des API implique l'utilisation de logiciels lui permettant de communiquer avec les serveurs supportant les API et de leur faire exécuter les fonctionnalités requises. Il n'existe pas véritablement de solution clef en main permettant de réaliser cet ensemble d'opérations simplement. On trouve majoritairement des modules logiciels jouant le rôle de connecteur fonctionnel avec un langage de programmation de plus ou moins haut niveau.

[3] Le plancher premium est fixé à environ 150$/mois et peut être multiplié par dix pour l'offre entreprise. Cette offre complète celle antérieure, beaucoup plus élevée, accessible auprès d'opérateurs chargés de la valorisation des flux de données.

L'investissement technique nécessaire pour les utiliser est faible lorsqu'il s'inscrit dans l'effort plus soutenu qu'implique la maîtrise des outils analytiques contemporains comme R par exemple.

Dans ce contexte, l'accès aux flux de données implique une démarche équivalente à celle d'un développeur logiciel qui doit d'abord déclarer son projet de développement auprès de la plateforme (Twitter) afin d'obtenir une clef d'authentification applicative (*consumer key*). Cette déclaration préalable est de nature contractuelle et engage le développeur à suivre les conditions générales de la politique de développement de Twitter[4]. Cette contractualisation vient en complément d'un autre engagement contractuel associé aux conditions générales d'usages (CGU) d'un compte Twitter. Il est fait référence à ce compte dans l'identification nécessaire via des clefs d'utilisation (*access token*) pour les services de Twitter. Ces différents contrats établissent la responsabilité personnelle (*accountability*) au regard de la balance entre les droits attribués et les devoirs contractés.

Parmi les devoirs du développeur figure le respect des modalités de sollicitation de l'API. Celles-ci font état des contrôles ascendants (client>serveur) ainsi que des contrôles descendants (serveur>client) suivant la nature des requêtes ou de l'état des services sollicités. Ces restrictions impactent notamment l'API SEARCH dont le fonctionnement est réglé en fréquence de requêtes et en volume de réponses. Pour l'API STREAM, les restrictions portent plutôt sur la complexité du filtre de flux.

Les différentes CGU définissent en partie les restrictions informationnelles opérées par Twitter. Au-delà de ces restrictions qui sont clairement énoncées et monitorées par la plateforme, d'autres contraintes encadrent les productions informationnelles disponibles aux bornes des API.

Les API RESTful

En informatique, le concept d'interface est associé à l'organisation logique des traitements. Les interfaces articulent des fonctionnalités qui se distinguent du fait de leur nature ou des modalités de leur mise en œuvre. Le principe des interfaces s'est imposé dans le génie logiciel comme le moyen de maîtriser la complexité et d'assurer des développements informatiques modulaires et indépendants.

L'acronyme REST (*REpresentational State Transfer*) qualifie des interfaces donnant accès à des représentations associées aux entités manipulées (ou objets) dans l'application, selon un cahier des charges spécifique tendant à standardiser la production. Les directives REST reprennent les principales caractéristiques de la définition du Web dont en particulier le protocole de communication HTTP pour le transport des données, l'architecture client-serveur et l'absence de mémorisation d'états de sessions. Outre l'uniformisation du Web, les objectifs associés à la norme REST portent sur la qualité des performances (côté serveur) quel que soit le dimensionnement (*scalability*) de la demande.

On associe aussi le terme REST à l'adjectif RESTful (paisible) pour caractériser la conformité des modalités d'accès aux données de l'application suivant ce standard. En effet, sa mise en œuvre favorise l'interopérabilité applicative et uniformise le travail de développement des programmeurs, l'allégeant du même coup. L'appellation RESTful est de ce fait revendiquée par les plateformes comme un gage de qualité sans pour autant se conformer entièrement au standard.

Dans le contexte du développement des plateformes de services, la caractérisation REST s'applique aux API développées, contribuant ainsi à l'émergence d'une norme d'usage. Le choix d'un développement de type REST permet de séparer clairement les rôles dans l'accomplissement du

. . . . . .

[4] https://developer.twitter.com/en/developer-terms/agreement-and-policy, consulté le 21/05/2018.

service en se rapportant au modèle client/serveur. En particulier, cela permet de déléguer au client le soin de contextualiser ses requêtes et de mémoriser les éléments de session pertinents, ce qui allège d'autant la charge du serveur.

Twitter comporte un ensemble d'API (plus d'une dizaine) dont l'étendue et les spécificités évoluent en fonction de l'offre de services et de la politique d'ouverture des données. L'abandon récent de la restriction des 140 caractères du corps de message pour aller vers un format documentaire multimédia plus étoffé n'est pas la moindre de ces évolutions. Dans les faits, la limitation est maintenue pour l'affichage du texte mais ne l'est plus pour l'encodage textuel du message. Les textes peuvent désormais être plus longs mais masqués. Les informations structurelles du message, comme par exemple les comptes mentionnés, ou les URLs tweetées ne sont plus comptabilisées. Cette définition étendue s'aligne avec celle du billet d'information enrichie imposée par la concurrence, notamment de Facebook. Elle vise également la facilité de réponse pour favoriser le régime conversationnel des échanges et en conséquence la production de contenus.

Cette modification affecte le nombre et la nature des objets de Twitter. La structure informationnelle correspondante s'enrichit notamment de relations structurelles inédites. Il en découle de nouvelles possibilités de représenter et d'accéder aux informations associées à ces entités. Ces possibilités se traduisent par des représentations structurées qui se redéploient suivant une organisation d'interfaces elle-même renouvelée. Toutefois, la rétrocompatibilité qui est nécessaire avec les développements antérieurs de la plateforme Twitter et de son écosystème atténue la portée de ces évolutions. Toutes ces API ne respectent pas strictement la définition REST mais la plupart d'entre elles en sont cependant dérivées.

Les représentations publiques gratuites

Compte tenu des enjeux portés par les données, nous distinguons formellement deux systèmes de représentations. Le premier répond aux logiques du système d'information qui supporte l'activité de la plateforme et ses développements stratégiques. Ce système de représentation (interne) est confidentiel. Même si une partie des données qu'il organise fait l'objet de publicisation, celle-ci mobilise un système de représentation (externe) distinct. Ce deuxième système répond aux nécessités d'une communication maîtrisée respectant les contraintes réglementaires, éthiques et stratégiques. Dans le cas de Twitter, la définition et l'organisation des API sont étroitement liées à ce système. En effet, la décomposition et les fonctionnalités des API reprennent les principes d'une conception centrée objet. Dans le cas présent, les représentations numériques mises en œuvre dans ces interfaces formalisent les entités conceptuelles et les actions associées aux modèles d'information et d'interaction adoptés par cette plateforme.

C'est ainsi que la représentation d'un tweet (STATUS) contient les représentations d'entités associées à l'auteur (USER), à des listes de composants (HASHTAG, MEDIA, URL, SYMBOL) et enfin, à l'extension du message qui correspond soit au tweet originel (STATUS) dans le cas d'un retweet soit à une liste de composants supplémentaires dans le cas d'un tweet long. L'attribut QUOTED_STATUS permet de repérer les cas de citations commentées, c'est-à-dire ajoutant une contribution à ce qui sinon n'aurait été qu'une rediffusion (RETWEETED_STATUS).

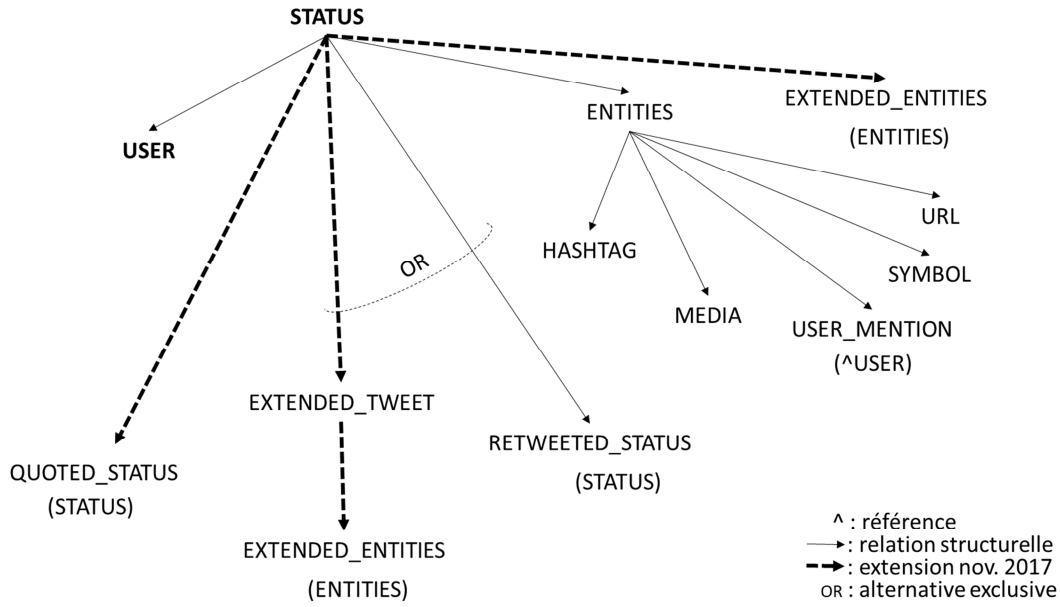


Figure 1. Schéma structurel externe d'un Tweet[5].

Seuls les deux représentations des STATUS et des USER comportent une clef d'identification interne qui en assure l'unicité et constitue le moyen de formuler des requêtes pour d'autres API connexes.

Les clefs d'identification fournies rendent compte de l'existence des deux entités STATUS et USER dans le système représentationnel. Leur présence conjointe dans la représentation du message rend probable l'existence formelle de la relation éditoriale d'attribution (*author*) dans le système de représentation interne. L'inscription du message originel dans le cas d'une rediffusion permet en outre de faire exister formellement deux relations orientées entre entités USER : une relation de rediffusion (retweet) entre auteurs (rediffusé et rediffusant) ainsi qu'une relation de mentionnement (*mention*) entre l'auteur et chacun des comptes identifiés dans le message.

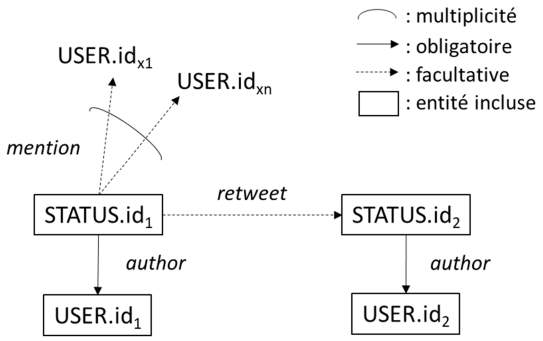


Figure 2. Schéma relationnel externe d'un Tweet

[5] La branche d'extensions (EXTENDED_TWEET) correspond à la sortie des restrictions de publications antérieures, en maintenant la cohérence avec le modèle historique.

Ces différents éléments déterminent l'ensemble des calculs que l'on peut opérer à partir de la capture d'un flux de Twitter. Il est ainsi formellement possible de rechercher dans ce flux et sans ambiguïté, l'ensemble des publications d'un abonné, d'établir l'existence de relations éditoriales entre auteurs et ainsi d'extrapoler des mesures d'intensité de flux ou de liaison entre ces acteurs, sous couvert d'hypothèses portant sur l'échantillonnage aux bornes des API. Cette unicité d'accès garantit en outre l'interopérabilité des services et des développements dans l'écosystème applicatif. Twitter privilégie ainsi l'exploitation des relations éditoriales mettant l'accent de cette manière sur une forme de réseautage que l'on peut qualifier de circonstanciel. Il s'agit en effet, de réseaux tributaires de la capture d'un flux qui ne vaut que dans le contexte spécifique de sa production et dans la durée des faits qu'il relate. L'analyse des graphes de rediffusion ne doit pas oublier qu'il s'agit de configurations dynamiques d'acteurs éventuellement éphémères. De ce point de vue, les représentations résultantes se distinguent de celles des réseaux déclaratifs de suivi (*follow*) dont la stabilité est plus grande mais dont le besoin de connaissance nécessite d'interroger une API distincte. Compte tenu des règles de cumul appliquées sur les sollicitations authentifiées des API, cette double interrogation ne peut pas être conduite simultanément avec une seule machine sans risque de discontinuité de flux. Ce jeu de contraintes techniques témoigne de la cohérence des limitations imposées et de l'éclatement fonctionnel entre les API de Twitter. Cette distribution informationnelle dans le système d'information publique a comme conséquence de réduire la possibilité d'analyse au fil de l'eau des réseaux d'acteurs et des flux informationnels. Twitter préserve ainsi, et pour son compte, la valeur événementielle du flux et l'offre de services qui en découle.

L'éditorialisation de flux

L'adaptation informationnelle réalisée par les opérateurs de plateformes du Web 2.0 s'inscrit dans une stratégie globale de mise en visibilité de leurs services et de contrôle des ressources disponibles. Elle constitue désormais ce que l'on peut considérer comme un processus d'éditorialisation (Lipsyc et Ihadjadene, 2013) généralisé à l'ensemble de ces ressources numériques. Ce processus s'applique en effet, non seulement à l'ensemble des publications sur le Web mais aussi dans les flux de données publicisés depuis les API. Le rôle de ces interfaces spécialisées n'est donc plus strictement de faciliter l'accès aux représentations numériques internes mais d'en contrôler l'accès, la diffusion et l'usage possible. La sélection des informations publicisées constitue la première étape de ce contrôle. La deuxième étape consiste à structurer le système représentationnel externe en fonction des objectifs de publicisation et des exigences de contrôle sur le potentiel heuristique des représentations produites.

L'alignement métaphorique

Pour Twitter, la logique de contrôle sur les contenus est cohérente avec un système métaphorique fondé sur les notions d'événement et de flux qui sous-tendent son fonctionnement. Dans ce contexte, la publication d'un message est assimilée à un événement éditorial. Ce dernier correspond à un événement qui survient puis s'estompe au regard de nouveaux événements plus « prégnants ». La prégnance d'un événement est définie par l'intensité relative de sa rediffusion (retweet) au sein de l'espace de publication courant. L'intérêt de ces métaphores est d'assurer la cohérence avec le modèle des médias de flux. Elles donnent un sens aux notions d'actualité (*news*) et de tendance (*trend*) qui concourent au succès de Twitter dans l'espace médiatique. Elles justifient également la fugacité des événements de publication et donc instaure la durée de vie des messages.

L'analogie avec les médias mise en œuvre par Twitter ne se limite pas aux conditions d'usages des services de publication proposés. Elle se prolonge dans la définition de ses API et des logiques de publicisation des données. À ce niveau, l'effacement progressif des données est réalisé dans le

système représentationnel au moyen d'un historique de publication auquel on ne peut accéder que de manière antéchronologique et pour un volume limité de messages. L'accumulation progressive de messages publiés dans l'historique conduit à un effacement d'autant plus rapide que le flux de publications est important2. Selon les modalités d'accès choisies, la portée historique d'une semaine (gratuit) est repoussée aux 30 derniers jours (premium) voire sans limite pour le mode « entreprise ».

Le mécanisme de rediffusion (retweet) est cohérent avec une communication de proche en proche portée par des réseaux de socialité. L'objectif associé est à la fois d'organiser la rareté de l'information et d'assurer la promotion d'un message en maintenant son actualité. Sa traduction fonctionnelle est d'assurer la rémanence des contenus informationnels du message originel dans l'historique de publications mais pour une durée qui ne dépasse que très rarement une semaine.

La mise en cohérence métaphorique naturalise en retour les restrictions d'accès aux données de publication. Le modèle éditorial des médias justifie, par exemple, l'identification de l'auteur originel d'un message rediffusé et non des intermédiaires qui ont participé à sa diffusion. Cette référence unique et légitime à la source est inspirée des bonnes pratiques de l'édition. Elle interdit en conséquence l'accès aux chaînes de rediffusion et de ce fait, à la connaissance des réseaux de diffusion effectifs (*followers*). De cette manière, l'identification des acteurs clefs de l'information est beaucoup plus difficile à appréhender, ce qui a des répercussions dans la mise en œuvre de modèles de communication.

Les réductions représentationnelles

D'autres formes de restrictions interviennent en complément des limitations de flux ; de nature représentationnelle, elles consistent à traduire sous une forme moins riche les représentations internes disponibles. Cet effacement n'est effectif qu'au niveau représentationnel externe. L'ensemble des messages (hors embargo) peuvent être retrouvés à partir de services payants assurés par des prestataires chargés de la monétisation des données de Twitter (Gnip, etc.). Nous distinguons principalement trois méthodes :

- le référencement revient à identifier les entités par un code interne plutôt que de fournir leurs représentations complètes. Il s'agit d'un moyen terme qui fait peser sur le client un coût supplémentaire (requête) s'il souhaite accéder à cette représentation sachant que l'interaction avec une API est aussi contrôlée. Dans tous les cas, la résolution référentielle désynchronise les représentations qui peuvent désigner deux états distincts d'une même entité. C'est le cas avec les comptes mentionnés mais aussi avec les URL qui pointent des éléments éventuellement externes au système de représentation interne de Twitter. Dès lors, les corrélations ne sont plus fondées temporellement ;
- l'écrasement consiste à restituer une forme de surface non structurée qui ne permet pas non plus d'établir une référence à une entité représentationnelle. Les composants d'un message (HASHTAG, etc.) rendent compte de limites qui tiennent autant au développement de la plateforme et de ses services qu'à la volonté de ne pas fournir les moyens d'une captation de valeur trop évidente. Si ces entités peuvent éventuellement avoir une existence formelle dans le système de représentation interne, elles n'en ont probablement pas dans le système externe. L'effort de représentation est ainsi reporté côté client et réduit la portée des services proposés.
- la synthèse est réalisée au moyen de catégories ou de quantifications produites à la place de représentations dénombrables d'entités. La quantification peut également s'accompagner d'un échelonnement des valeurs qui fixe une granularité de l'information fournie. Il est à noter ici que Twitter donne la valeur exacte des compteurs (*followers*, etc.) à la différence de

LinkedIn qui ne fournit pas de manière précise le nombre de relations personnelles au-delà de 500 contacts.

La distribution fonctionnelle

Comme nous avons pu le voir avec Twitter, les opérations précédentes s'appliquent sur une sélection d'attributs déterminée par le choix d'une distribution fonctionnelle. Il s'agit de faire en sorte qu'aucun appariement de collections de données ne permette, en quantité et en qualité, de produire une information estimée stratégique pour l'opérateur de la plateforme. Pour empêcher les appariements (ou croisements de données), il suffit que les représentations soient disjointes, c'est-à-dire qu'il n'y ait pas d'identifiant unique permettant une mise en correspondance immédiate. Si toutefois la jointure est inévitable, celle-ci doit alors avoir une portée limitée ou induire un coût de réalisation dissuasif ou monétisable.

Dans le cas de Twitter, ce contrôle global est réalisé en monitorant les transactions authentifiées. Le contrôle en fréquence et en volume de transactions allonge les délais de calcul, atténuant du même coup leur portée.

Conclusion

Certes moins privative que d'autres plateformes (Facebook, LinkedIn, etc.), l'évolution de la politique de publicisation des données de Twitter témoigne d'une économie numérique où la valeur s'établit désormais sur l'abondance des données.

Pour les plateformes de service, la publicisation des données a pour objectif de constituer un écosystème et de créer des opportunités de marchés. Cette ouverture des données ne peut s'envisager sans contrôle ni contrepartie de la valeur ainsi cédée. La mise en œuvre de modalités payantes contribue à cette régulation. Le ticket d'entrée élevé qui en découle établit de nouvelles frontières financières entre les tiers d'une économie numérique. On identifie donc, d'un côté les acteurs dotés de financements solides, porteurs de projets et disposant des moyens d'une exploitation massive de flux de données et de l'autre, des acteurs à la recherche d'opportunités, engagés dans une démarche exploratoire. Cette partition vaut également pour le monde académique qui ne fait pas exception, malgré l'évolution des cadres légaux[6]. Dans un contexte règlementaire prioritairement favorable à l'économie numérique, contourner l'écueil des coûts impose des regroupements institutionnels afin d'assumer le cofinancement d'études ou d'atteindre la masse critique nécessaire pour être éligible lors d'appels à projets. Le changement d'échelle imposé dans la conduite de tels projets fait écho aux processus de rationalisation porté par les institutions scientifiques. Toutefois, les efforts de coordination scientifique restent d'un effet limité puisqu'au regard des CGU fixées par Twitter, il ne peut être question d'ouvrir[7] des collections de Tweets sans s'exposer à des poursuites. Relâcher cette contrainte devient une nécessité scientifique et citoyenne pour laquelle nos communautés scientifiques doivent se mobiliser.

En dépit des restrictions qualitatives ou quantitatives opérées sur les données, les API maintiennent les caractéristiques de flux instantanés massifs, témoignant d'activités et d'interactions connectées à

. . . . . .

[6] https://www.economie.gouv.fr/republique-numerique, https://ec.europa.eu/research/openscience/index.cfm?pg=openaccess

[7] Au sens de l'open data que promeut la loi sur la république numérique. La seule solution jugée satisfaisante par Twitter étant de ne partager que les identifiants de Tweets (Status.id) pour une période inférieure à 30 jours. Pour plus de détail voir le post : https://blog.ldodds.com/2017/05/19/can-you-publish-tweets-as-open-data/ (consulté le 01-06-2018)

l'espace social et à la sphère médiatique. Ainsi, l'accès aux flux de Twitter est une source d'intérêt pour l'étude de phénomènes info-communicationnels, dans la limite bien comprise des mécanismes de publicisation et de leurs conséquences sur les données. La compréhension qualitative des données disponibles est une première étape d'un processus analytique mobilisant des techniques et des méthodes mixtes, à la fois qualitatives et quantitatives. Elle offre également la possibilité d'une lecture critique de l'économie de l'information sous-jacente à ce type d'entreprise du Web. Enfin, le changement d'échelle, conduisant aux données massives constitue un objectif distinct que l'on peut néanmoins situer dans la continuité de cette acculturation progressive aux données de l'Internet.

Références bibliographiques

Bigot, Jean-Edouard, Julliard, Virginie, et Mabi, Clément, (2016), « Humanités numériques et analyse des controverses au regard des SIC. Retour sur une expérience pédagogique », *Revue française des sciences de l'information et de la communication*, n°8.

Bouquillion, Philippe, et Matthews, Jacob Thomas, (2010), *Le Web collaboratif. Mutations des industries de la culture et de la communication*, Grenoble : Presses universitaires de Grenoble.

Institut Montaigne (2015), *Big data et objets connectés - faire de la France un champion de la révolution numérique*, Rapport. En ligne : http://www.institutmontaigne.org/res/files/publications/rapport%20objets%20connecte%CC%81s.pdf consulté le 1 juillet 2018.

Lipsyc, Carole, et Ihadjadène, Madjid, (2013), « Architecture de l'information et éditorialisation. In L'architecture de l'information : un concept opératoire », *Études de communication*, (2), n°41, p. 103-118.

Morstatter, Fred, Pfeffer, Jürgen, et Liu Huan, (2014), « When is it Biased? Assessing the Representativeness of Twitter's Streaming API », in *Proceeding WWW '14 Companion Proceedings of the 23rd International Conference on World Wide Web*. Seoul, Korea p. 555-556.

O'Reilly, Tim, (2005), « What is Web 2.0? », in *Online communication and Collaboration* - A Reader Ed. Donelan H., Kear K. And Ramage M. The Open University - Routledge p. 225-234.

Paquienséguy, Françoise et al. (2017), « Manifeste pour un positionnement des sciences de l'information communication (SIC) vis-à-vis des Digital Studies (DS) et autres mutations du Numérique », Sous la direction de Françoise Paquienséguy. *Revue française des sciences de l'information et de la communication*, n°10.

Les médiations de l'*open data* au prisme des applications liées à la mobilité

The mediations of open data through the prism of mobility applications

Las mediaciones de los datos abiertos a través de las aplicaciones dedicadas a la movilidad

Article inédit, mis en ligne le 15 novembre 2018.

Valentyna Dymytrova

Valentyna Dymytrova est chercheure post doctorante en Sciences de l'Information et de la Communication au laboratoire EA 4147 Elico. Dans le cadre de l'ANR-14-CE24-0029 OpenSensingCity (2014-2018), ses recherches portent sur les dispositifs, les acteurs et les discours de l'open data.

valentyna.dymytrova@sciencespo-lyon.fr

Plan de l'article

Résumé

Cet article étudie les médiations de la donnée et les processus de traduction dont elles font l'objet au prisme des applications liées à la mobilité. À partir d'une méthodologie qualitative, nous analysons des entretiens avec des acteurs de l'*open data* et de la *smart city* et des discours d'accompagnement des applications. Les services urbains assurés par des applications résultent des négociations et des alliances entre producteurs de données, collectivités territoriales, entreprises des télécoms et du logiciel, entrepreneurs du web, constructeurs de terminaux et usagers. Dans cette fabrique de la ville, la figure de l'usager-citadin est tiraillée entre un équipement technologique lui promettant une implication à travers une rationalisation et une individualisation des services et des infrastructures largement automatisées qui font de lui un usager discipliné.

Mots clés

Open data, application, médiation, mobilité, *smart city*, ANT.

Abstract

This article studies the mediations of the data and the translation processes through the prism of applications related to mobility. Using a case study methodology, we analyze interviews with open data and smart city actors and promotional discourses related to applications. Application-driven urban services result from negotiations and alliances between data producers, local authorities, telecom and software companies, web entrepreneurs, terminal builders and users. In this factory of the city, the figure of the user-city is torn between technological equipment promising him an implication through a rationalization and an individualization of the services and largely automated infrastructures which make him a disciplined user.

Keywords

Open data, application, mediation, mobility, smart city, ANT.

Resumen

Este artículo estudia las mediaciones de los datos y los procesos de traducción a través del prisma de las aplicaciones dedicadas a la movilidad. Usando una metodología de estudio de casos, analizamos entrevistas con actores de datos abiertos y ciudades inteligentes y discursos promocionales de esas aplicaciones. Los servicios urbanos impulsados por la aplicación son el resultado de negociaciones y alianzas entre productores de datos, autoridades locales, empresas de software y telecomunicaciones, emprendedores web, constructores de terminales y usuarios. En esta fábrica de la ciudad, la figura de la ciudad-usuario se debate entre el equipamiento tecnológico que le promete una implicación a través de una racionalización y una individualización de los servicios y las infraestructuras en gran parte automatizadas que lo convierten en un usuario disciplinado.

Palabras clave

Open data, aplicaciones, mediación, movilidad, *smart city*, ANT.

Introduction

L'open data désigne aujourd'hui à la fois une injonction à la mise à disposition des données administratives ou issues du secteur public et un processus d'ouverture progressive des données publiques qui se traduit généralement par leur publication via des catalogues et des portails dédiés (Chignard, 2012). La mise à disposition de données publiques donne l'impression que la « gouvernementalité algorithmique » (Rouvroy, Berns, 2010) n'est plus le monopole de la puissance publique et des grandes organisations privées.

Si l'open data permet *a priori* à chacun de réutiliser librement les données publiques, la conception des portails de l'open data et les modes d'accès aux données qu'ils proposent ne sont pas à la portée des citoyens sans compétence technique (Paquienséguy, 2016). Les discours d'accompagnement de l'ouverture des données publiques considèrent que les citoyens *lambda* auront accès aux données par l'intermédiaire d'applications et de services, développés par des professionnels des data et des

« pro-am »[1] (Dymytrova, Paquienséguy, 2017). Dans ce cas-là, les données ouvertes libérées sont intégrées dans de nouvelles « boites noires » au risque de renforcer les pouvoirs existants (Goëta, 2015).

Contrairement au nouveau « régime de vérité numérique », fondé sur « l'utopie d'un accès immédiat au réel comme tel » et à l'idéologie des Big data considérant « la donnée comme un fait ultime, parlant d'elle-même sans médiation » (Rouvroy, 2014), la publication et la réutilisation des données ouvertes sont marquées par une prolifération des médiations de la donnée (Goëta, 2015). Qu'elles soient informatiques, info-communicationnelles ou pédagogiques (Larroche, Vila, 2015), ces médiations consistent en un ensemble d'actions cherchant à transformer les données en informations. Si les médiations informatiques se matérialisent dans des applications informatiques comme par exemple des plateformes mettant à disposition des données ouvertes, les médiations pédagogiques recouvrent les fonctions de sensibilisation et de vulgarisation des données. Elles peuvent faire partie des applications, mais sont surtout assurées par des associations comme par exemple FING et OpenDataFrance. Quant aux médiations info-communicationnelles, elles consistent dans la manipulation de données sélectionnées dans le but de fournir un service à visée potentiellement marchande. Dans ce cadre, les données subissent des traitements informatiques qui les rendent lisibles et compréhensibles par des non experts. Les médiations info-communicationnelles se présentent par exemple sous forme d'informations, fournies par des applications, qui constituent la partie visible de l'open data pour le grand public.

Les applications mobiles incarnent une nouvelle forme de services au cœur des modèles de la ville intelligente, soumise aux impératifs d'attractivité et de qualité dans un contexte de compétition interurbaine (Picon, 2013). Notion polysémique, la *smart city* renvoie au passage de l'infrastructure aux services à travers l'usage généralisé de données numériques urbaines. Si le développement des villes dites aujourd'hui « numériques » s'inscrit historiquement en France dans les politiques de décentralisation et de régionalisation des années 1980, il se poursuit avec la généralisation d'Internet et des réseaux de télécommunication intégrant les infotechnologies dans l'aménagement du territoire (Loechel, 2000). Les dispositifs mobiles et ultra-mobiles (ordinateurs, portables, téléphones mobiles, tablettes), les réseaux mobiles (GSM, 4G, réseaux sans fil, Bluetooth) et les technologies de l'information sont aujourd'hui considérés comme un vecteur puissant d'innovation urbaine.

En prenant appui sur des modèles relatifs à la sociologie de l'acteur réseau ou ANT (Actor-Network Theory) (Akrich, Callon, Latour, 2006), nous analyserons les modalités de médiation de la donnée, qui concernent les applications liées à la mobilité. Plus précisément, nous étudions comment les applications sont produites (acteurs, moyens matériels, technologiques et symboliques) et comment elles re-contextualisent les données qu'elles utilisent pour produire de nouveaux services urbains (traitements et modalités d'interprétations). L'appareillage conceptuel de l'ANT permettra d'établir un ensemble de « points de vue » sur la production des applications et de caractériser les modalités de médiation de la donnée en fonction des processus de traduction, assurés par des actants humains et non humains (Callon, 2006).

Notre terrain s'appuie sur une enquête menée en France de février 2016 à avril 2017 dans le cadre de l'ANR OpenSensingCity 14-CE24-0029[2]. Celle-ci étudiait les conditions de production et de réutilisation des données ouvertes en France (Dymytrova et al., 2017). Pour cet article, nous mobilisons les entretiens semi-directifs d'une durée de 45 minutes environ, menés du février 2016 au

. . . . . .

[1] Le terme désigne la figure du professionnel-amateur : celui qui a le statut d'amateur mais qui s'investit dans son activité avec les mêmes exigences et les mêmes règles de qualité que les professionnels.

[2] L'équipe Elico est dirigée par F. Paquienséguy et composée des chercheurs suivants : V. Dymytrova, I. Hare, V. Larroche, M-F. Pereylong et M. Vila.

février 2017 à Paris, Lyon, Nantes, Toulouse et Grenoble, auprès de différents acteurs du domaine de l'open data et de la *smart city* (n=45). Le panel des personnes interrogées dans des structures publiques et privées comprend des développeurs d'applications, des producteurs des données de la mobilité, des chefs de projet open data, des fournisseurs des plateformes et des analystes des données de la mobilité. Ces matériaux sont complétés par une analyse des discours d'accompagnement des applications liées à la mobilité, disponibles dans Google Play Store et l'App Store d'Apple.

Nos analyses s'inscrivent dans la méthodologie de l'étude de cas, définie comme une enquête empirique étudiant un phénomène contemporain dont les frontières ne sont pas suffisamment nettes et qui pose des questions du type « comment » ou « pourquoi » (Yin, 2002). Nous articulons ainsi une analyse thématique des entretiens et une analyse des discours promotionnels qui accompagnent les applications liées à la mobilité dans les magasins d'application.

Après avoir introduit brièvement notre modèle théorique, nous analyserons comment les concepteurs inscrivent leurs « visions du monde » dans les applications liées à la mobilité et les discours qui les accompagnent. Ensuite, nous présenterons les médiations de la donnée en lien avec des acteurs qui les assurent. Enfin, nous montrerons en quoi le dispositif d'application est un réseau sociotechnique, constitué des interactions entre les actants humains et non humains, en nous focalisant sur des processus de traduction autour des médiations de la donnée.

Les dispositifs d'application sous l'angle de l'ANT

L'ANT invite à explorer des traces des interactions, « négociations et ajustements entre actants humains et non humains » (Akrich, Callon, Latour, 2006, p. 271), par lesquelles ils établissent et maintiennent des connexions entre eux et échangent des propriétés. Comme les acteurs ont un ensemble diversifié d'intérêts, la stabilité du réseau repose principalement sur les alignements des acteurs grâce aux « traductions » (Callon, 1990 ; Law, 1992), c'est-à-dire leur capacité de réinterpréter ou d'approprier les intérêts d'autrui, et aux « intéressements », c'est-à-dire stratégies d'alliance, de négociation et de mise en connexion d'acteurs issus de plusieurs univers. Les traductions vont de pair avec « l'inscription » qui consiste à traduire une intention ou un intérêt « incorporés dans les textes, les machines, les compétences corporelles » en un support matériel (Callon, 1990, p. 143).

À l'instar de tout acteur-réseau, le dispositif d'application est un réseau sociotechnique, composé d'un ensemble d'actants hétérogènes, humains et non humains, reliés entre eux par des associations particulières. En effet, les applications liées à la mobilité reposent sur un large réseau sociotechnique qui rassemble quatre mondes : le monde des acteurs des transports qui ont une expertise dans le domaine de la mobilité et produisent des données dans le cadre des délégations de service public, celui des collectivités territoriales qui gèrent l'ouverture des données et souhaitent le développement économique de leur territoire grâce aux réutilisations des données, celui des entrepreneurs du web qui utilisent les données pour les transformer en services et enfin, celui des usagers qui s'approprient des informations *via* des applications et qui enrichissent les services proposés avec leurs propres données, fournies de manière volontaire ou automatiquement générées par des objets connectés utilisés.

Les applications liées à la mobilité, nouveaux services au sein de la *smart city*

L'engouement autour de l'open data a coïncidé avec l'arrivée sur le marché des smartphones et des applications sur l'AppStore d'Apple en 2007 (Chignard, 2012). De fait, les concours d'application pour un usage mobile deviennent une forme fréquente d'accompagnement des démarches de l'open

data par des administrations et des collectivités (Goëta, 2015). Considérées comme particulièrement porteuses de valeurs d'usage, les données de la mobilité sont parmi les plus réutilisées lors des concours[3] : « *Aujourd'hui, c'est la locomotive qui tire les wagons, on le voit avec la SNCF, la RATP, les applications comme Citymapper ou d'autres, ce sont les données faciles, elles sont parlantes. Les horaires de transport, les retards, ça parle à beaucoup de gens, analyser un PLU ou un budget municipal, c'est beaucoup plus difficile* » (entretien avec un chef de projet open data d'une métropole).

Une large partie du travail des concepteurs consiste à « inscrire » (Akrich, 2006) leur vision du monde dans le contenu technique du nouvel objet. Les idées et les scenarii des applications liées à la mobilité viennent souvent des intuitions des développeurs qui sont eux aussi des usagers, confrontés à une variété de pratiques de déplacement. Par exemple, l'idée de l'application Géovélo développée par la Compagnie des Mobilités pour une dizaine de villes françaises a émergé dans le cadre de l'association de promotion du vélo qui était amenée à faire beaucoup de cartes papier pour aider les cyclistes dans leurs déplacements : « *Le but c'est d'améliorer la prise en compte des transports doux dans les trajets quotidiens des gens. Ce n'est pas forcément très mis en valeur par les applications des opérateurs qui se borment en général à faire du calcul d'itinéraires de manière très pauvre* » (entretien avec un développeur créateur de Géovélo).

Dans le cadre des concours, comme par exemple celui pour la mobilité en Isère, lancé en 2017 par le département d'Isère et la société Cityway, la thématique et les jeux de donnes mis à disposition ont considérablement délimité des scenarii éventuels. Toutefois, les développeurs ont avancé un certain nombre d'hypothèses sur le « monde » qu'ils souhaitaient inscrire dans leur application. Ils ont « traduit » sur le plan technique à la fois les contraintes du concours et leurs aspirations d'« améliorer le quotidien de la population et de compléter les applications disponibles fournies par les opérateurs » (entretien avec un développeur créateur de Mobili.watch, lauréat du concours).

Si le nombre d'applications dédiées à la mobilité est considérable, les discours qui les accompagnent développent plusieurs promesses dont les plus récurrentes sont le gain de temps, la gestion d'incertitude et le service personnalisé. Par exemple, l'application Moovit, conçue par une *start-up* israélienne, s'est imposée comme acteur important du secteur de la mobilité pour plus de 2000 villes à l'échelle internationale. Elle se positionne comme « votre assistant personnel pour les transports en commun » en promettant « le meilleur itinéraire possible » et le « déplacement sans stress » grâce au système d'alertes en cas de perturbations du trafic. Dans un autre domaine, l'application Géovélo propose une personnalisation de parcours des cyclistes en fonction de type de vélos, de la vitesse moyenne et du parcours souhaité pour « assurer confort, sécurité et tranquillité ». Dans le domaine du stationnement, ParkingMap, développé par une *start-up* parisienne pour plusieurs métropoles françaises promet de faire « gagner du temps » et de rendre « la ville plus facile », grâce aux différentes offres de stationnement en temps réel.

Les discours d'accompagnement expriment ainsi les façons dont les concepteurs cherchent à intéresser les usagers en « traduisant » leurs attentes en fonction des impératifs de rationalisation (efficacité/rapidité) et d'individualisation/personnalisation. En effet, les usagers jouent un rôle important dans le succès d'une application, en l'expérimentant, en l'évaluant et en favorisant sa diffusion (notes, commentaires). Ils contribuent aussi au fonctionnement d'une application en l'enrichissant avec leurs retours d'expérience et en laissant les traces de leurs déplacements.

. . . . . .

[3]La mise à disposition de ces données est encouragée par la directive européenne INSPIRE, 2007 et récemment la législation nationale Loi Macron, 2015 et Loi Lemaire, 2016.

La conception de l'application : quelles médiations des données ?

La conception d'une application comprend une série de médiations des données, notamment des médiations informatiques et des médiations info-communicationnelles (Fig.1). Les deux sont interdépendantes et assurées par des actants humains et non humains que nous pouvons qualifier de « médiateurs », car ils « transforment, traduisent, distordent et modifient le sens ou les éléments qu'ils sont censés transporter » (Latour, 2007, p.58).

D'abord, les développeurs récupèrent et agrègent toutes les données, ouvertes ou non qui sont nécessaires à la réalisation d'un scénario. L'accès aux données et aux contextes de leur production conditionnent la capacité à les interpréter et à leur apporter de la valeur. Car rares sont les données qui ont été produites en fonction des usages prévus par une application. Une fois les données récupérées et agrégées, les développeurs les intègrent dans une base de données interne qui permet de les traiter et de les gérer.

Médiations	Médiateurs	Opérations	Acteurs
Médiations info-communicationnelles	Éditeurs d'applications mobiles	Réutiliser la donnée	Développeurs, pro-ams, *start-ups*, grandes entreprises...
	Marchés d'applications	Assurer la visibilité et l'accès aux données *via* l'application	Opérateur téléphoniques, éditeurs de systèmes d'exploitation, constructeurs de terminaux mobiles...
	Usagers récepteurs et producteurs de contenus	Produire des traces, des flux et des contributions	Individus, groupes sociaux, citoyens...
Médiations informatiques	Producteurs des données	Produire des données	Délégataires de services publics, entreprises spécialisées, citoyens...
	Diffuseurs des données	Publier des données	Collectivités, administrations, entreprises...
	Portails et plateformes OD	Définir les modalités d'accès aux données	Collectivités, administrations, entreprises...
	Éditeurs de logiciels et de plateformes	Définir les modalités d'accès aux données	Sociétés privées

Figure 1. Modalités des médiations de la donnée.

Les traitements constituent une partie importante dans la réutilisation des données car elles permettent aux développeurs d'« inscrire» leur vision du monde dans le dispositif d'application. Les données ouvertes sont réutilisées par des développeurs dans un nouveau contexte pour répondre à

un problème particulier ce qui fait de chaque application un objet unique, même si les données sources sont identiques.

En fonction des utilisations prévues, les traitements comprennent des filtrages, des nettoyages, des consolidations et des transformations des données dans des formats compatibles avec les solutions utilisées. En effet, les développeurs doivent « traduire » leurs objectifs dans les contenus techniques en tenant compte des configurations des appareils mobiles auxquels ils destinent leur application. Par exemple, certains services proposés par les applications comme le guidage dans l'espace dépendent des équipements des terminaux avec une caméra, des modes de géolocalisation (GPS, Wifi, antenne-relais, IP) et des outils de détection d'orientation de l'appareil (accéléromètre, gyromètre, magnétomètre).

En même temps, les applications doivent être développées dans un respect plus ou moins strict des recommandations des constructeurs des terminaux mobiles qui gèrent leur validation et leur distribution *via* des espaces numériques dédiés, appelés marchés d'applications, dont les plus importants sont actuellement Apple Store, Android Google Play et Microsoft Windows Phone. Chaque constructeur préconise non seulement l'usage de langages de programmation spécifiques mais propose également des modules et des composants d'interface prédéfinis qui peuvent être directement exploitables ou adaptés aux projets des développeurs.

L'étape finale de la production d'une application consiste en une vérification des données et leur intégration dans les couches logicielles du « front office » de l'application. Là encore, nous sommes face à une médiation qui assurerait le passage des résultats du travail d'un intégrateur de données et d'un développeur « back office » au travail d'un développeur « front office ».

La forme et le contenu des applications résultent ainsi d'une double médiation, à la fois technique et sociale qui articule différents acteurs et champs sociaux.

Les données au cœur des processus de traduction

À la fois des constructions techniques et des constructions sociales, les applications tissent des relations et des interactions plus ou moins fortes entre des entités hétérogènes qui rassemblent des actants humains (les producteurs de données, les collectivités qui les diffusent, les développeurs qui les réutilisent et les usagers finaux) et des actants non humains (les jeux de données, les plateformes et portails OD, les langages de programmation, les smartphones). La figure 2 représente d'une manière non exhaustive[4] les actants impliqués dans ce réseau sociotechnique et pointe quelques processus de traduction qui assurent son fonctionnement et sa stabilité.

[4] Nous ne pouvons aborder ici ni des acteurs d'intermédiation comme infolabs ou fablabs qui font le lien entre les différents acteurs de l'écosystème de l'innovation, ni des technologies et des langages de programmation qui permettent de traiter les données et de les intégrer aux bases de données internes aux applications.

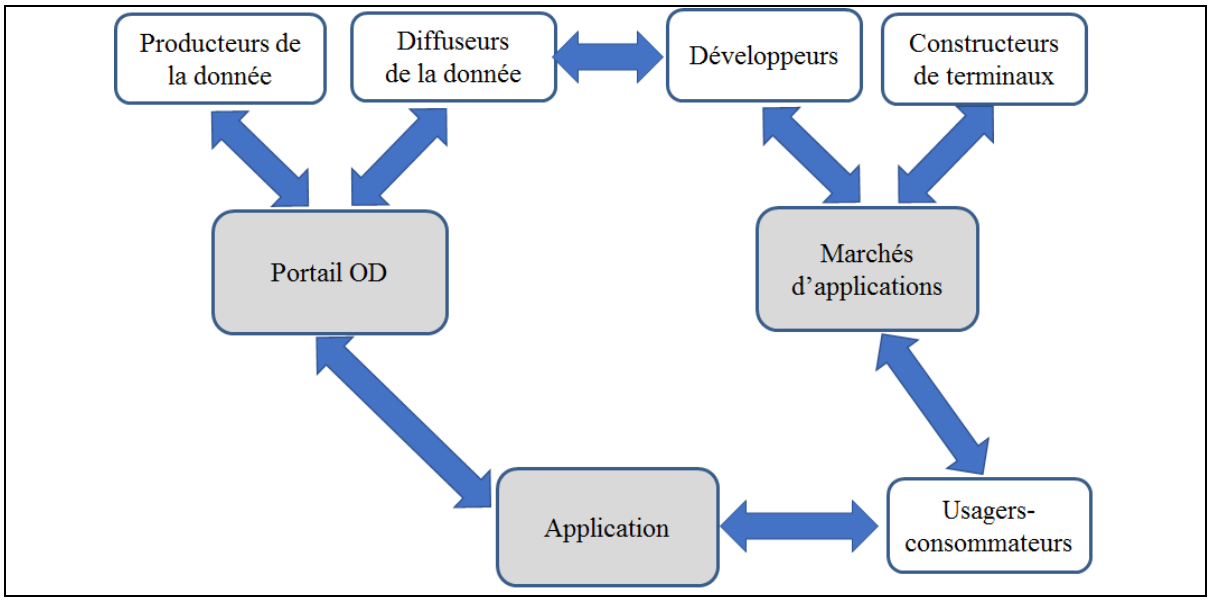


Figure 2. Réseau sociotechnique du dispositif d'application.

Les applications fonctionnent à partir des données, qu'elles soient ouvertes ou propriétaires. Les collectivités produisent rarement les données liées à la mobilité. Elles les récupèrent généralement en négociant en amont avec les entreprises qui les génèrent dans le cadre de la délégation de services publics.

Initialement saisies dans les formats métiers[5] à travers des applications métiers bien particuliers, les données sont collectées, générées et agrégées par des producteurs dans une logique qui leur est propre. Elles sont ainsi constituées d' « attachements » à des métiers, à des systèmes d'information, à des applications, à des opérateurs et à des modèles économiques ou juridiques (Courmont, 2016). La législation, les politiques publiques et les stratégies des collectivités territoriales obligent les producteurs à «instaurer » les données qu'ils mettent à disposition.

Une fois identifiées et extraites, les données sont rigoureusement retraitées ou « brutifiées » (Denis, Goëta, 2013) pour passer des données métiers à des données considérées comme ouvertes à différentes réutilisations et publiables sur les plateformes dédiées.

Les diffuseurs publient les données sur des portails et plateformes dédiés en définissant des modalités d'accès à ces données dans leurs aspects techniques, économiques et juridiques. Inscrits dans les enjeux politiques et territoriaux, les plateformes et les portails open data industrialisent l'accès aux données en les rendant disponibles et directement exploitables.

La conception de la plateforme, ses règles de fonctionnement et ses modalités de gouvernance matérialisent les choix des administrations qui les ont conçues et font partie intégrante des stratégies d'« intéressement » qui cherchent à intégrer les développeurs dans les plateformes (Paquienséguy, 2016). En effet, les plateformes de l'open data facilitent considérablement le travail des développeurs : « Les données officielles, statiques des plateformes, c'est toujours plus simple pour nous, parce qu'elles sont de meilleure qualité » (entretien avec un développeur).

. . . . . .

[5] Il s'agit d'un format relevant d'usage habituel dans un domaine professionnel. Par exemple, les délibérations du conseil municipal sont souvent en PDF, des plans d'urbanisme ou des cadastres en DWG. Dans le domaine des transports, ce sont des formats Neptune, porté par les professionnels de la mobilité et GTFS, conçu par Google et porté par les entrepreneurs du web.

Les données relatives à la mobilité sont souvent mises à disposition sous forme d'API (Application Programming Interface), un ensemble de fonctions logicielles qui peuvent être appelées depuis l'extérieur de l'application qui les expose. Les API offrent un accès sélectif aux informations et non à l'intégralité du fichier en téléchargement ce qui permet d'interagir facilement avec les données en les recherchant et les récupérant automatiquement. De fait, les API présupposent un certain type d'usages des données : « Dans une API, les fichiers sont cachés, on est obligé de passer par un serveur et le serveur peut filtrer ou poser des limites. Par exemple, l'API de la SNCF est limitée à 1000 appels par jour. Cela oblige à surveiller sa consommation des données et à négocier avec les responsables des données » (entretien avec un développeur et responsable technique d'une *start-up*).

À l'heure actuelle, la plupart des métropoles[6] construit sa plateforme open data et ses jeux de données en fonction de ses propres choix techniques et politiques (Paquienséguy, Dymytrova, 2017). Il en résulte des « plateformes maison », sans mise en cohérence des jeux de données, de leurs structurations et de leurs formats entre les plateformes des différentes collectivités : « Chacun met à disposition des données qu'il possède. De fait, la plupart des applications liées à la mobilité sont souvent limitées à un territoire ».

Comme nous l'avons déjà souligné, chaque constructeur non seulement dote son terminal d'outils pour se repérer dans l'espace qui peuvent être intégrés aux fonctionnalités des applications mais définit aussi son environnement technologique fondé sur des langages de programmation spécifiques et propose des modules et des composants d'interface prédéfinis qui peuvent être directement exploitables ou adaptés aux objectifs des développeurs. En plus, les constructeurs de terminaux mobiles gèrent la validation et la distribution des applications *via* des marchés d'application. Les développeurs sont amenés à composer avec des prescriptions pour des terminaux mobiles auxquels ils destinent leur application.

Enfin, les usagers-consommateurs constituent un actant important du réseau sociotechnique analysé. Ils sont en lien à la fois avec les marchés d'applications où ils vont choisir et télécharger une application en fonction de leurs attentes et des traductions de celles-ci à travers les discours qui accompagnent et promeuvent les applications. Ils sont également en lien avec les applications car ils contribuent d'une manière volontaire et involontaire à nourrir les applications utilisées avec leurs données.

En effet, les applications croisent généralement l'open data avec d'autres sources de données, comme les données privées, les données collaboratives partagées *via* des plateformes ouvertes de type OpenStreetMap (par ex., Géovélo), les données collectées par différents capteurs (par ex., ParkingMap) ou les données crowdsourcées, générées par les usagers des applications. Un véritable « troc de données » se met en place entre les applications et les collectivités : « Je vous donne mes données et en échange vous me donnez des infos que vous avez, sur les travaux, sur les boucles de comptage, etc... pour enrichir ma propre donnée » (entretien avec un Country Manager de l'application Moovit).

Chacun des éléments cités contribue, à sa manière, au fonctionnement de l'application. En même temps, cette activité collective du réseau sociotechnique constitue une « boite noire », invisible pour l'usager et qui s'ouvre brutalement lors des échecs, des incidents et des pannes. Par exemple, les surcharges du serveur, une limitation d'usage d'une API en temps ou en nombre d'appels ou encore une évolution des modes de captation, de formats ou des modalités d'exploitation des données

. . . . . .

[6]La loi n° 2016-1321 du 7 octobre 2016 pour une République numérique oblige les collectivités dont la population est supérieure à 3 500 habitants de publier leurs données dans un format ouvert et exploitable, soit sur leur propre plateforme open data, soit sur la plateforme nationale data.gouv.fr.

menacent le bon fonctionnement d'une application tout en remettant en question les enrôlements et les alliances composant l'acteur-réseau.

Conclusion

Dans cet article, nous avons cherché à montrer que les informations telles qu'elles se présentent au citoyen-usager sur l'écran de son objet connecté ne sont pas l'expression directe des données métiers mais font l'objet d'une série de médiations des données qui reposent sur un large réseau sociotechnique. En effet, les nouveaux services urbains qui reposent sur la circulation et sur la valorisation des données sont issus des négociations et des alliances entre des collectivités territoriales, des grandes entreprises issues des secteurs des télécoms et du logiciel, des *start-ups* et des constructeurs de terminaux. La forme et le contenu des applications résultent d'une double médiation technique et sociale et reflètent des orientations des acteurs dominants qui en maîtrisent la production et la distribution. La médiation des données se présente alors comme un processus d'articulation, d'interpénétration et d'interaction entre les champs sociaux, qui « doit compter avec des équilibres fluctuants de tensions et de forces entre les différents éléments » (Floris, 1995, p.147). Nous assistons ainsi à une industrialisation des données publiques selon un modèle de « convergence à plusieurs facettes » qui « met en jeu de multiples catégories d'acteurs dont les stratégies ne convergent pas nécessairement et dont les « intérêts » peuvent même s'opposer durablement » (Miège, 2006, p.175). Si les appropriations sociétales des applications restent à étudier, l'analyse du processus concret de la production du dispositif d'application repose la question des rapports de force entre différents acteurs de l'open data dans la fabrique de la ville face à « l'informationnalisation » (Miège, 2004) croissante de la société où la figure de l'usager-citadin est tiraillée entre un équipement technologique promettant une implication des citoyens à travers une rationalisation et une individualisation des services et des infrastructures largement automatisées qui font de lui un usager passif, voire discipliné.

Références bibliographiques

Akrich, Madeleine (2006), « La description des objets techniques » (p. 159-178), in Akrich, Madeleine, Callon Michel, Latour, Bruno, *Sociologie de la traduction – Textes fondateurs*, Paris : Presses de l'École des Mines.

Akrich, Madeleine, Callon Michel, Latour, Bruno (2006), *Sociologie de la traduction – Textes fondateurs*, Paris : Presses de l'École des Mines.

Callon, M. (1990), "Techno-economic Networks and Irreversibility", *The Sociological Review*, vol. 38, n° 1, p. 132-161, [en ligne], consulté le 27 février 2018, http://cast.b-ap.net/arc590s14/wp-content/uploads/sites/28/2014/04/Callon_Techno-economic-networks-and-irreversibility.pdf.

Callon, Michel (2006), « Sociologie de l'acteur réseau » (p. 267-276), in Akrich, Madeleine, Callon Michel, Latour, Bruno, *Sociologie de la traduction – Textes fondateurs*, Paris : Presses de l'École des Mines.

Chignard, Simon (2012), Open data : Comprendre l'ouverture des données publiques, Paris : FYP Editions.

Courmont, Antoine (2016), *Politiques des données urbaines : ce que l'open data fait au gouvernement urbain*, thèse de doctorat en sciences politiques, Sciences Po Paris, [en ligne], consulté le 2 février 2018, http://spire.sciencespo.fr/hdl:/2441/4014o907at8roohlckove1dov3.

Denis, Jérôme, Goëta, Samuel (2016), *"Brutification" et instauration des données. La fabrique attentionnée de l'open data*, [en ligne], consulté le 27 février 2018, http://i3.cnrs.fr/workingpaper/brutification-et-instauration-des-donnees-la-fabrique-attentionnee-de-lopen-data/.

Dymytrova, Valentyna, Larroche, Valérie, Paquienséguy, Françoise, Peyrelong, Marie-France (2017), « Open Data et Smart Cities : Quels chantiers pour les SIC ? », *Les Cahiers de la SFSIC*, n°14, juin 2017, p. 308-313.

Dymytrova, Valentyna, Paquienséguy, Françoise (2017), « La réutilisation et les réutilisateurs des données ouvertes en France : une approche centrée sur les usagers », *Revue Internationale des Gouvernements Ouverts*, vol. 5, p. 117-132, [en ligne], consulté le 27 février 2018, http://ojs.imodev.org/index.php/RIGO/article/view/204/338.

Goëta, Samuel (2015), « L'open data : une forme ultime de transparence ? » (p. 49-67), in Catellani, Andrea ; Crucifix, Audrey ; Hambursin, Christine ; Libaert, Thierry (dir.), *La communication transparente*, Louvain : Presses universitaires de Louvain.

Floris, Bernard (1995), « Les médiations dans les rapports sociaux », *Réseaux*, vol. 69, n° 1, p. 141-156.

Larroche, Valérie, Vila, Martine (2015), « Urban Data et stratégies dans le secteur des services : Le cas de la métropole lyonnaise » (p. 183-197), in Broudoux, Evelyne ; Chartron , Ghislaine, *Big data - Open data: Quelles valeurs ? Quels enjeux ?*, Louvain-la-Neuve : De Boeck Supérieur.

Latour, Bruno (2007), *Changer de société, refaire de la sociologie*, Paris : La Découverte.

Loechel, André (2000), « Naissance et développement des villes numériques », *Les cahiers du numérique*, vol.1, n°1, p. 15-55.

Miège, Bernard (2006), « Les industries culturelles et médiatiques : une approche socio-économique » (p. 163-180), in Olivesi Stéphane (dir.), *Sciences de l'information et de la communication : Objets, savoirs, discipline*, Presses universitaires de Grenoble.

Miège, Bernard (2004), *L'information-communication, objet de connaissance.* Bruxelles : De Boeck & Larcier ; Bry-sur-Marne : INA.

Paquienséguy, Françoise (2016), « Smart city & Open Data : à qui profitent les données ouvertes ? », *colloque international des Sciences du Territoires 2016*, Proceedings « En quête de territoire (s), Looking for territories », Grenoble, 17-18 mars 2016, p. 351-357.

Paquienséguy, Françoise, Dymytrova, Valentyna (2017), *Analyse de portails métropolitains de données ouvertes à l'échelle internationale*, livrable 2, Elico, [en ligne], consulté le 26 février 2018, hal-01449348.

Picon, Antoine (2013), Smart cities. Théorie et critique d'un idéal auto-réalisateur, Paris : Éditions B2.

Rouvroy, Antoinette, Berns, Thomas (2010), « Le nouveau pouvoir statistique: Ou quand le contrôle s'exerce sur un réel normé, docile et sans événement car constitué de corps "numériques" », *Multitudes*, n°40, p. 88-103.

Rouvroy, Antoinette (2014), « Des données sans personne: le fétichisme de la donnée à caractère personnel à l'épreuve de l'idéologie des Big Data », contribution en marge de l'Étude annuelle du Conseil d'État. *Le numérique et les droits et libertés fondamentaux*, [en ligne], consulté le 20 février 2018, http://works.bepress.com/antoinette_rouvroy/55.

Yin, Robert (2009), *Case Study Research. Designs and Methods*, Newbury Park: Sage.

Les machines apprenantes et la (re)production de la société : les enjeux communicationnels de la socialisation algorithmique

The learning machines and the (re)production of the society: the communicational stakes of algorithmic socialization

Máquinas de aprendizaje y (re)producción en la sociedad: los retos de comunicación de la socialización algorítmica

Article inédit, mis en ligne le 15 novembre 2018.

Jean-Sébastien Vayre

Jean-Sébastien Vayre a réalisé une thèse de doctorat à l'université Toulouse Jean Jaurès (CERTOP – UMR 5044). Il est actuellement post-doctorant au LITEM (EA 7363) et est rattaché à l'Institut Mines-Télécom Business School. Ses travaux portent sur le développement du big data et de l'intelligence artificielle au sein des organisations et des marchés. jean-sebastien.vayre@imt-bs.eu.

Plan de l'article

Résumé

Comment les technologies d'intelligence artificielle, en automatisant la communication organisationnelle, peuvent-elles contribuer à la (re)production de la société ? Afin de traiter cette question, nous commencerons par exposer ce que nous entendons par socialisation algorithmique. Nous montrerons ensuite que, d'un point de vue théorique, les technologies d'apprentissage artificiel ont pour but d'automatiser les processus de communication au sein des organisations selon des mécanismes de socialisation algorithmique qui favorisent la reproduction de la société. Puis nous soutiendrons que, d'un point de vue empirique, ces machines apprenantes recouvrent un important travail de cadrage de leurs activités inférentielles. Nous verrons de cette façon que l'automatisation de

la communication qu'autorisent ces technologies renvoie *in fine* à des formes de socialisation algorithmique qui contribuent plus à la production qu'à la reproduction de la société.

Mots clés

Socialisation algorithmique, production et reproduction sociale, intelligence et apprentissage artificiels, mégadonnées.

Abstract

How can artificial intelligence technologies, by automating organizational communication, contribute to society's (re)production? In order to address this question, we will start by explaining what we mean by algorithmic socialization. We will then show that, from a theoretical point of view, artificial learning technologies aim to automate communication processes within organizations according to algorithmic socialization mechanisms that promote the reproduction of society. Then we will argue that, from an empirical point of view, these learning machines cover an important work of framing their inferential activities. We will see in this way that the automation of communication that these technologies allow ultimately refers to forms of algorithmic socialization that contribute more to the production than to the reproduction of society.

Keywords

Algorithmic socialization, social production and reproduction, artificial intelligence and learning, big data.

Resumen

¿Cómo pueden las tecnologías de inteligencia artificial, al automatizar la comunicación organizativa, contribuir a la (re)producción de la sociedad? Para responder a esta pregunta, comenzaremos explicando lo que entendemos por socialización algorítmica. A continuación mostraremos que, desde un punto de vista teórico, las tecnologías artificiales de aprendizaje apuntan a automatizar los procesos de comunicación dentro de las organizaciones según mecanismos algorítmicos de socialización que promueven la reproducción de la sociedad. Entonces argumentaremos que, desde un punto de vista empírico, estas máquinas de aprendizaje cubren un trabajo importante de enmarcar sus actividades inferenciales. Veremos así que la automatización de la comunicación que permiten estas tecnologías se refiere en última instancia a formas de socialización algorítmica que contribuyen más a la producción que a la reproducción de la sociedad.

Palabras clave

Socialización algorítmica, producción y reproducción social, inteligencia y aprendizaje artificiales, mega datos.

Introduction

Comme le montre les travaux de Pierre Musso (2017), les sociétés occidentales contemporaines sont caractérisées par une triple dynamique : la révolution managériale, le développement de la cybernétique et celui des Technologies de l'Information et de la Communication (TIC). Plusieurs auteurs ont ainsi mis en avant que les évolutions récemment élaborées dans le domaine de l'apprentissage artificiel[1] (i.e. *machine learning*) sont au cœur du mouvement de numérisation que connaissent actuellement les organisations (Boullier, 2016 ; Cardon, 2015 ; Ganascia, 2017). En permettant aux machines d'apprendre à partir de grandes masses de données, ces évolutions doivent autoriser l'automatisation de la communication d'une part toujours plus importante des informations que les hommes et les machines s'échangent pour coordonner leurs actions, et faire exister leurs organisations. L'objectif est alors de favoriser la modularisation des procédés de production et de distribution des biens de consommation (Maistre, 2016 ; Laumond, 2016) de manière à préparer l'avènement d'une société de services véritablement personnalisés (Kohler & Weisz, 2016).

Les algorithmes d'apprentissage que les acteurs économiques conçoivent et implémentent au sein de leurs entreprises ont de cette façon pour but d'accroître l'« agentivité » (i.e., *agency* ; Pickering, 1995) des systèmes d'information qui les composent. Ils doivent permettre aux machines d'inférer des connaissances sur le monde de façon à ce qu'elles puissent communiquer aux entités sociales et techniques qui forment l'organisation les informations lui permettant de s'adapter aux évolutions de son environnement. Aussi, à l'ère de ce que les professionnels appellent la société 4.0, la question du rôle des machines est centrale : les enjeux communicationnels qui sont associés à l'accroissement progressif de l'autonomie cognitive que leur confèrent les humains sont importants. C'est pourquoi nous proposons dans cet article de traiter le questionnement suivant : en automatisant la communication organisationnelle, comment les technologies d'intelligence artificielle qui sont aujourd'hui développées peuvent-elles contribuer à la (re)production de la société ?

Pour traiter cette problématique, nous commencerons par exposer les cadres théorique et méthodologique que nous avons mobilisés pour effectuer notre enquête. Ceci nous amènera à présenter la notion de socialisation algorithmique ainsi que les limites de notre étude qui porte avant tout sur le cas des technologies d'apprentissage artificiel appliquées à la gestion de la relation client. Nous exposerons ensuite nos résultats en deux grandes sections. Dans la première, nous verrons que, sur le plan théorique, les technologies d'apprentissage artificiel ont pour but d'automatiser des processus de communication selon des mécanismes de socialisation algorithmique qui favorisent la reproduction de la société. Dans la seconde section, nous soutiendrons que, sur le plan empirique, les machines apprenantes auxquelles nous nous sommes intéressé recouvrent un important travail de cadrage de leurs activités inférentielles. Nous verrons de cette manière que, non plus d'un point de vue abstrait mais concret, l'automatisation de la communication qu'autorisent ces technologies renvoie *in fine* à des formes de socialisation algorithmique qui contribuent plus à la production qu'à la reproduction de la société.

. . . . . .

[1] Nous mobilisons la notion d'apprentissage artificiel plutôt que celle d'apprentissage automatique dans la mesure où, comme le soulignent Antoine Cornuéjols et Laurent Miclet (2010), le qualificatif « automatique » est ambigu. Dans les domaines des sciences cognitives et de l'informatique, il renvoie en effet à des formes de traitement de l'information dit de bas niveau. Or, les technologies d'apprentissage artificiel peuvent réaliser des traitements de haut niveau.

Cadres théorique et méthodologique

Dans cette première section, nous présentons les cadres théorique et méthodologique que nous avons déployés pour réaliser notre enquête. Cette partie nous permettra donc d'exposer notre positionnement disciplinaire qui se trouve au croisement des sciences de l'information, de la communication, de la sociologie des sciences et de celle des techniques. En référence aux travaux de Bernard Miège (2007), l'étude que nous présentons dans cet article relève des sciences de l'information et de la communication puisqu'elle consiste à examiner : « *des processus d'information et de communication relevant d'actions contextualisées, finalisées, prenant appui sur des techniques, sur des dispositifs, et participant des médiations sociales et culturelles* »[2] (Miège, 2007, p. 199-200). Comme nous l'avons souligné en introduction, les technologies d'apprentissage artificiel ont pour fonction d'automatiser une part toujours plus grande des processus de communication qui se déroulent entre l'immense diversité des acteurs sociaux qui font usages des TIC. Aussi, nos travaux relèvent des sociologies des sciences et des techniques dans le sens où ils portent un intérêt particulier aux rôles que jouent ces technologies d'apprentissage artificiel au sein des sociétés occidentales contemporaines.

Afin d'exposer les cadres théorique et méthodologique que nous avons élaborés pour effectuer notre étude, nous proposons de commencer par présenter le concept d'agentivité et la manière dont il peut être compris dans le cas des objets techniques. Nous soulignerons de cette façon l'intérêt de la notion de socialisation algorithmique pour saisir les enjeux communicationnels associés au développement des machines apprenantes. Nous exposerons ensuite les matériaux sur lesquels s'appuie notre enquête. Et nous pointerons les limites de notre étude qui porte essentiellement sur le cas des technologies d'apprentissage artificiel appliquées à la gestion de la relation client.

Cadre théorique

La notion d'agentivité a beaucoup été utilisée dans le domaine de la psychologie sociale afin de pointer les capacités d'action dont les humains disposent pour organiser leur environnement (Jézégou, 2014). Ce concept a plus exactement été forgé en opposition aux travaux des sociologues les plus déterministes qui pensent qu'au cours de leurs socialisations, les individus incorporent les normes, les règles et les valeurs qui composent la société. Par exemple, à la différence du sociologue Pierre Bourdieu (1970), le psychologue Albert Bandura (1986) ne considère pas les humains comme des agents se conformant passivement aux rôles que leur impose le collectif. Il les appréhende plutôt comme des acteurs qui savent anticiper et ajuster leurs actions en fonction des croyances qu'ils ont de leurs compétences, de celles des autrui et des objectifs qui orientent leurs activités. Dans le sens des travaux de Georges H. Mead (2015), la notion d'agentivité permet par là même d'insister sur les capacités d'autodirection dont disposent les humains pour construire leur socialisation.

Autrement dit, le fait de mobiliser ou non le concept d'agentivité constitue un bon indicateur de la façon dont les chercheurs en sciences humaines et sociales se représentent la société (Giddens, 1986). Avec cette notion, l'ordre social n'est plus structuré par les forces mystérieuses d'un être surplombant qui s'appellerait « société ». Il est la résultante des interactions qu'entretiennent les humains : c'est par le biais de l'agentivité que les individus sont capables de produire et de communiquer les informations leur permettant de préserver ou de transformer les normes, les valeurs et les règles de la société de manière à construire son historicité. C'est donc à travers cette même agentivité que les humains peuvent contribuer à la reproduction de la société – c'est-à-dire

[2] Précisons que, selon Bernard Miège (2007), cette citation est extraite du texte rédigé par le Conseil National Universitaire (CNU) pour définir ce qui est du ressort de la section 71.

maintenir sa stabilité dans la durée (*cf.* la notion de statique sociale) – mais aussi à sa production – c'est-à-dire orienter son évolution dans le temps (*cf.* la notion de dynamique sociale).

Dans la continuité des travaux susmentionnés, de nombreux auteurs se sont appropriés la notion d'agentivité pour désigner, non plus les capacités d'action des humains sur le monde, mais celles des objets techniques (Bouillon, 2015 ; Callon, 1998 ; Cooren & Fairhurst, 2009 ; Denis & Pontille, 2010 ; Latour, 1994). Car ceux-ci sont porteurs de scénarii d'usage qui sont élaborés à travers diverses activités de « description » des mondes à l'intérieur desquels ils doivent être implémentés et d'« inscription » de ces mondes dans la conception même de ces objets (Akrich, 1987 ; 1989). Pour Madeleine Akrich (1987 ; 1989), les objets techniques peuvent de ce fait être compris comme des signes porteurs de sens dans la mesure où ils communiquent plus ou moins explicitement les visions de ceux qui les conçoivent : les objets techniques recouvrent des schèmes cognitifs qui, une fois cristallisés dans leurs conceptions même, renvoient à des « prescriptions » d'usage qui doivent être correctement interprétées par les utilisateurs (Akrich, 2004). Et c'est précisément en ce sens que ces objets sont dotés d'une agentivité : à l'instar des humains, ils participent à la structuration de la société, c'est-à-dire à sa (re)production.

À l'heure actuelle, un tel élargissement du concept d'agentivité est pertinent pour au moins une raison. Les applications socioéconomiques des technologies d'apprentissage artificiel ont pour finalité de conférer une certaine autonomie cognitive aux machines de façon à ce qu'elles puissent communiquer aux entités humaines et matérielles qui composent l'organisation des informations leur permettant d'adapter leurs comportements en fonction : d'une part, des évolutions de leurs environnements et, d'autre part, des objectifs que leurs concepteurs leur prêtent (Vayre, 2016). Or, si l'on accepte le fait qu'il puisse exister une certaine symétrie entre les capacités d'action des hommes et des techniques, il apparaît que le problème que pose l'apprentissage autonome des machines fait écho à celui que pose la socialisation des humains, à tout le moins selon la conception que certains d'entre d'eux s'en font. Rappelons par exemple que, dans la théorie de l'action de Talcott Parsons (2005), ce problème peut être formulé comme suit : comment une organisation peut-elle faire en sorte que les individus qui la composent s'adaptent à son fonctionnement (*Adaptation*) et suivent ses objectifs (*Goal-attainment*) de façon à garantir leurs intégrations au sein du groupe (*Integration*) et, par voie de conséquence, la stabilité du collectif (*Latency*) ? C'est donc ici que se trouve l'intérêt du concept de socialisation algorithmique : désigner la principale fonction sociotechnique des technologies d'apprentissage artificiel. Car c'est par le biais de cette socialisation d'un type particulier que les humains doivent conférer aux machines l'autonomie cognitive leur permettant de communiquer aux entités sociotechniques qui forment l'organisation les informations autorisant son adaptation à l'environnement que représentent les données qui les nourrissent (i.e., les *big data*). Et cela, sans pour autant limiter la capacité de contrôle des concepteurs de ces technologies, c'est-à-dire la possibilité, pour ces derniers, d'encadrer et d'orienter les apprentissages qu'elles réalisent.

Partant, en référence aux travaux de Yves Jeanneret (2008) sur la trivialité et par le biais de la notion de socialisation algorithmique, nous souhaitons montrer que les machines apprenantes participent à faire exister les dimensions logistique, sociale et poétique des processus de communication organisationnelle. Afin de permettre l'automatisation de ces processus, ces machines organisent la combinaison et la circulation des traces numériques que recouvrent les *big data* (*cf.* la dimension logistique) en fonction des jeux de pouvoir qui associent leurs concepteurs, leurs implémenteurs et leurs utilisateurs (*cf.* la dimension sociale). Par conséquent et contrairement à ce que certains spécialistes proclament, les informations que ces technologies produisent et communiquent ne sont pas de strictes reproductions des régularités comportementales que les *big data* doivent permettre d'identifier : les machines apprenantes n'atteignent pas le réel de manière immanente. Les informations qu'elles fabriquent et transmettent au sein des organisations doivent plutôt être comprises comme des créations dans la mesure où elles participent *in fine* à la transformation des

entités cognitives qui forment la culture d'une communauté (*cf.* la dimension poétique). C'est donc en ce sens que nous proposons, dans cet article, de tester l'hypothèse exploratoire (He) qui suit :

> **Hypothèse exploratoire (He)**
>
> De façon analogue aux humains, les technologies d'apprentissage artificiel que ceux-ci implémentent au sein de leurs organisations recouvrent des mécanismes de socialisation qui, au sens stricte du terme, produisent la société plus qu'ils ne la reproduisent.

Cadre méthodologique

Pour mettre à l'épreuve cette hypothèse, nous nous appuierons sur différents matériaux recueillis au cours d'une enquête qui s'est déroulée du mois de septembre 2012 au mois de septembre 2015.

Le premier matériau est constitué d'entretiens et d'observations que nous avons effectués de 2013 à 2015 auprès de cinq entreprises développant des technologies d'intelligence et d'apprentissage artificiels appliquées, pour la plupart d'entre elles, au secteur de la gestion de la relation client (*cf.* tableau 1). Ces entretiens et ces observations avaient pour finalité de nous permettre de mieux comprendre la conception, le fonctionnement et l'implémentation de ces technologies au sein des organisations commerciales. Ils ont été conduits durant cinq projets de collaboration dont deux ont donné lieu à des partenariats de 18 mois chacun.

Identifiant	Statut	Activités
S_01	Société par action simplifiée qui a été active pendant 3 ans. Son effectif était d'un peu moins de 10 salariés. Son chiffre d'affaire en 2013 était d'environ 100 000 €. Elle a été fermée en 2016	L'entreprise S_01 était spécialisée dans la conception d'algorithmes prédictifs appliqués à la recommandation de biens de consommation
S_02	Société par action simplifiée qui est en activité depuis 6 ans. Son effectif est d'un peu plus de 20 salariés. Son chiffre d'affaire en 2013 était de plus de 300 000 €. Elle recense 1 établissement actif	L'entreprise S_02 est spécialisée dans la conception de logiciels de personnalisation des environnements numériques marchands qui reposent sur différentes technologies de prédiction
S_03	Société par action simplifiée qui est en activité depuis 7 ans. Son effectif est d'un peu plus de 20 salariés. Son chiffre d'affaire en 2017 est de plus de 1 000 000 €. Elle recense 2 établissements actifs	L'entreprise S_03 est spécialisée dans la conception de logiciels de personnalisation des environnements numériques marchands qui reposent sur différentes technologies de prédiction
S_04	Société par action simplifiée qui est en activité depuis 8 ans. Son effectif est d'un peu plus de 10 salariés. Son chiffre d'affaire en 2016 est de plus de 600 000 €. Elle recense 1 établissement actif	L'entreprise S_04 est spécialisée dans la conception d'un moteur de recherche dit sémantique et reposant sur un principe d'analyse dit contextuel
S_05	Société par action simplifiée à associé unique qui est en activité depuis 10 ans. En 2012, son effectif est d'un peu plus de 20 salariés. Il est de 0 salarié en 2018. Son chiffre d'affaire en 2012 est de plus de 3 000 000 €. Elle recense 1 établissement actif	L'entreprise S_05 est spécialisée dans la conception d'un annuaire de « bonnes adresses » personnalisé disponible sous forme de site Internet et d'application mobile

Tableau 1 : Présentation des sociétés enquêtées

Le troisième matériau d'enquête est composé d'entretiens qui ont été réalisés avec treize professionnels des données (i.e., *data scientists*) et qui ont été conduits sur le thème des activités de conception des technologies d'apprentissage artificiel. Le tableau 2 expose les statuts et les formations de l'ensemble de ces professionnels qui, lorsqu'ils ne sont pas spécialisés dans le domaine de la gestion de la relation client ont, au cours de leur carrière, plusieurs fois eu l'occasion de travailler dans ce secteur d'activité.

Identifiant	Statut	Formation
E_01	Co-fondateur d'une entreprise développant des logiciels de gestion innovants	CentraleSupélec, Master en sciences + Université Paris XI, DEA en traitement du signal
E_02	Fondatrice d'une entreprise de traitement et d'analyse de mégadonnées	INSA Toulouse, Master d'ingénieur en mathématiques et informatique appliqués
E_03	Fondateur d'une entreprise développant des logiciels de traitement de données textuelles + consultant dans une société de services informatiques spécialisée dans le traitement et l'analyse des mégadonnées	ENSAE ParisTech, Formation en sciences des données + ENS Cachan, Master en apprentissage statistique + Université Paris I, Doctorat en mathématiques et informatique appliqués
E_04	Ingénieur spécialisé dans le traitement et l'analyse de mégadonnées au sein d'une des plus importantes entreprise du secteur du numérique	CentraleSupélec, Master en mathématiques et informatique appliquées
E_05	Directeur marketing d'une entreprise spécialisée dans la diffusion numérique de la presse	INSAI, Master d'ingénieur en statistique et marketing
E_06	Directeur des données d'une banque spécialisée dans le secteur touristique et immobilier	Ecole polytechnique, Formation en informatique, cryptographie et réseaux + ENS Cachan, Master en apprentissage statistique + ENPC, Formation en mathématiques et informatique pour l'ingénieur
E_07	Adjointe en chef d'un des bureaux de la direction générale du trésor	Ecole Polytechnique, Formation d'ingénieur + ENSAE, Formation en statistique et économie
E_08	Consultant dans une société de services informatiques spécialisée dans le traitement et l'analyse de mégadonnées	Université Paris VI, Master en intelligence artificielle
E_09	Directeur des données d'une entreprise développant des jeux vidéo dits sociaux	ISUP, Master en actuariat, fouille de données et apprentissage statistique
E_10	Fondateur d'une entreprise développant des logiciels de personnalisation des environnements numériques marchands	ENS Cachan, Master en mathématiques + Université Paris VII, Doctorat en mathématiques
E_11	Fondateur d'une entreprise développant un moteur de recherche dit de nouvelle génération	ISTEG, Master en commerce et marketing
E_12	Fondateur d'une entreprise éditrice de plateforme de formation	IUT de Créteil, Formation en génie électrique option informatique
E_13	Directeur département automobile d'un institut d'études et de conseils	Université de Bordeaux I, Master en sciences économiques et gestion

Tableau 2 : Présentation des statuts et des formations des professionnels interviewés

La socialisation algorithmique et la reproduction de la société

Comme nous l'avons annoncé en introduction, cette deuxième section a pour objet de soutenir l'hypothèse selon laquelle, sur le plan purement théorique, les technologies d'apprentissage artificiel appliquées à la gestion de la relation client automatisent des processus de communication organisationnelle selon des mécanismes de socialisation algorithmique qui favorisent la reproduction de la société. Nous nous appuierons pour cela sur l'exemple des réseaux de neurones artificiels qui connaissent un grand succès auprès des acteurs économiques engagés dans le développement du numérique. Rappelons toutefois que, dans la troisième section, nous discuterons cette première hypothèse en soutenant cette fois-ci que, sur le plan empirique, les technologies d'apprentissage artificiel que nous avons étudiées encouragent, au final, la production de la société.

L'exemple des réseaux de neurones artificiels…

Du point de vue de notre problématique, les réseaux de neurones artificiels forment un cas d'étude intéressant pour au moins deux raisons. La première est qu'ils sont construits selon le modèle du perceptron qui est une des plus anciennes technologies d'apprentissage artificiel (McCulloch & Pitts,

1943 ; Rosenblatt, 1958). La deuxième raison est que, comme le montre leur grand succès auprès de *Google*, *Amazon*, ou encore, *Facebook*, ces technologies sont aujourd'hui très appréciées par les professionnels qui sont engagés dans la numérisation des organisations et des marchés. Une des principales causes de cet engouement est qu'avec l'augmentation des puissances de calcul des ordinateurs au cours de ces dernières années, les réseaux de neurones peuvent actuellement être implémentés au sein de petits systèmes informatiques (e.g. un téléphone) de façon à permettre leur adaptation automatique à leur environnement d'usage[3]. Cette adaptation est alors permise par un procédé d'apprentissage extrêmement simple qui est plus connu sous le nom de règle de Hebb et qui veut qu'un peu à la manière des neurones biologiques, plus les unités de calcul d'un perceptron s'activent en même temps et plus elles sont liées les unes aux autres (Hebb, 1949). À partir de la formule exposée dans la partie droite de la figure 1, le réseau de neurones artificiels représenté dans la partie gauche pourrait ainsi apprendre à distinguer, par le biais de la sortie *y*, si l'image représentée par les pixels x_1, x_2, ..., x_n forme ou non un carré.

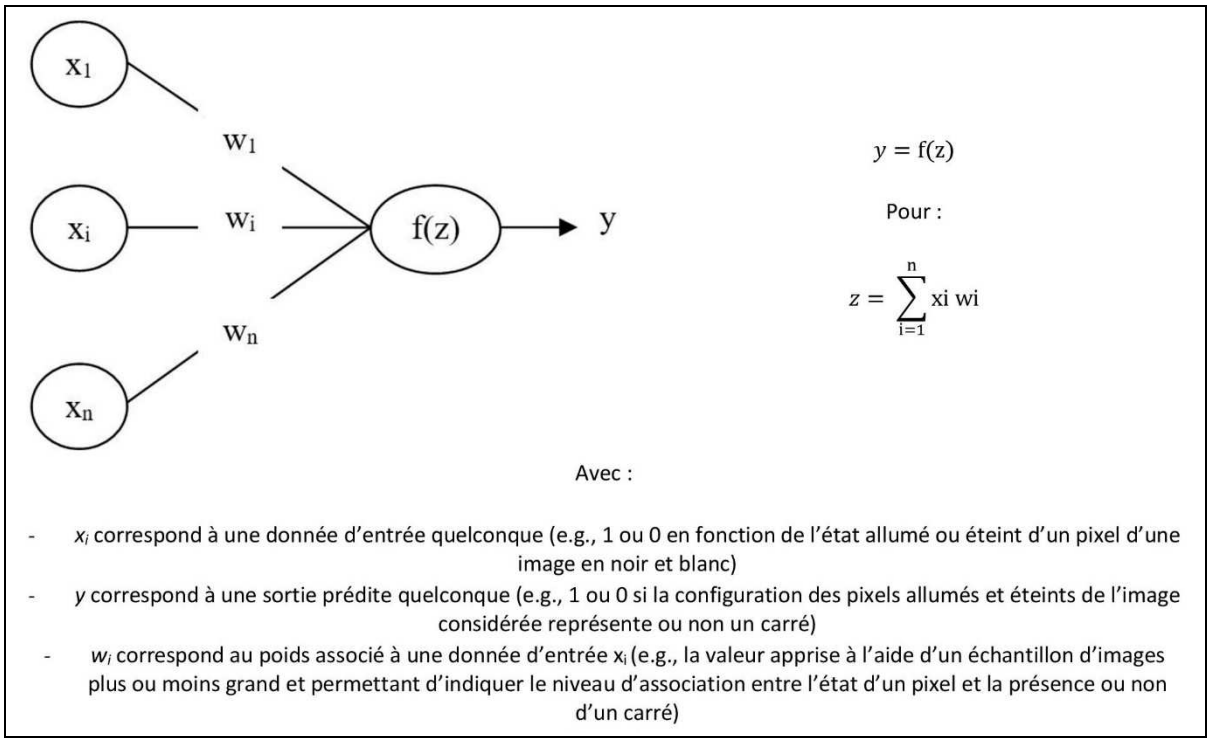


Figure 1 : Réseau de neurones artificiels

Notons alors que cet apprentissage pourrait être effectué au fil de l'eau (*cf.* Cornuéjols & Miclet, 2010). Par exemple, la technologie d'intelligence artificielle élaborée par la société S_3 s'apparente à un réseau de neurones profond qui est capable d'apprendre les préférences des consommateurs à partir :

. . . . . .

[3] C'est par exemple le cas de l'application *SwiftKey* qui est conçue à partir d'un réseau de neurones artificiels capable d'apprendre les habitudes d'écriture des utilisateurs des *smartphones* de manière à faciliter la rédaction de texte. Précisons toutefois qu'à l'heure actuelle, il reste toujours difficile de faire fonctionner les réseaux de neurones les plus complexes sur les supports mobiles.

- des données d'entrée x_i que sont les attributs des produits que composent leurs fiches de présentation et les traces d'usage que forment les données permettant de renseigner la machine sur les comportements des consommateurs ;
- des données de sortie y_i que sont les indices de préférence que composent les notations que ces mêmes consommateurs confèrent à ces biens, l'attention qu'ils leur accordent en termes de durée de consultation et l'intérêt qu'ils leur portent en les plaçant ou non dans leurs paniers.

En référence aux travaux de Dominique Cardon et d'Antonio Casilli (2015), la large diffusion des technologies d'apprentissage artificiel dans le domaine de la relation client participe ainsi à redéfinir les frontières entre travail et consommation. Il ne faut en effet jamais oublier que c'est par le biais des indices de préférence produits par les consommateurs, que les machines peuvent effectuer les apprentissages leur permettant de communiquer à ces mêmes clients des informations qui doivent leur apparaître comme pertinentes, et cela, de façon à produire de la plus-value économique pour les marchands. À l'instar des dispositifs de gestion étudiés par Marie-Anne Dujarier (2008) ou par Guillaume Tiffon (2013), le système d'intelligence artificielle élaboré par la société S_3 est un cas caractéristique : selon un processus de communication automatique similaire à celui que nous venons d'exposer, son architecture cognitive est conçue pour « mettre au travail » les consommateurs en économicisant une partie de leurs activités de recherche d'information sur les sites de vente-à-distance. D'après nous, ce point mérite d'être souligné, mais aussi d'être mieux considéré. Il nous rappelle en effet que toute technologie d'apprentissage artificiel, pour fonctionner, nécessite beaucoup de travail humain pour organiser l'environnement que forment les bases de données qui les nourrissent ; même si ce travail est parfois imperceptible compte tenu de son caractère automatique.

… et la reproduction des biais de représentation

Comme l'indique la citation 1 et le *verbatim* 1 exposés dans la figure 2, les technologies d'apprentissage artificiel que sont les réseaux de neurones artificiels comporteraient deux grands avantages par rapport aux technologies d'intelligence artificielle plus traditionnelles que sont, par exemple, les systèmes experts. Premièrement, à la différence de XCON qui constitue, rappelons-le, un des plus anciens systèmes experts (Crevier, 1997 ; Forgy, 1981), les réseaux de neurones artificiels n'ont pas besoin que leurs concepteurs définissent les règles d'inférence leur permettant de décider, par exemple, des configurations d'ordinateurs dont ont besoin tel et tel groupes de clients. À condition de disposer d'une quantité de données suffisante, ces technologies sont capables d'apprendre ces règles par elles-mêmes. Autrement dit, elles sont capables de découvrir et de coder seules les connaissances dont elles ont besoin pour réaliser leur travail inférentiel, qui pourrait donc consister, par exemple, à apparier des configurations d'ordinateurs et des clients. Comme le montre le cas de XCON (*cf.* citation 1), cette autonomie cognitive est importante pour les acteurs socioéconomiques dans la mesure où la production de biens et de services personnalisés implique une tolérance de la complexité qui manquait cruellement aux systèmes experts.

Citation 1 : La première version de *XCON* fut livrée à la *DEC* en janvier 1980. Le système n'était alors capable de configurer qu'un seul type de *VAX*, le *11/780*, mais, l'essai s'étant révélé concluant, la *DEC* dona son accord pour qu'il soit graduellement étendu à l'ensemble de sa production. En 1984, *XCON* était devenu le symbole de la réussite des systèmes experts : doté de trois cents règles en 1979, il en composait alors dix fois plus, trois mille, et servait à configurer dix modèles d'ordinateurs différents (p. 193) […]. La *DEC* n'avait eu de cesse de lui ajouter de nouvelles règles pour le maintenir en phase avec les constants changements et améliorations de la chaîne de production et, en 1987, *XCON* avait enflé au point de renfermer quelque dix mille règles, ce qui ne le fit pas gagner en efficacité, mais le rendit simplement plus gros. La *DEC* dut bientôt débourser deux millions de dollars par an pour mettre à jour la base de connaissances, et la plaisanterie courut dans la firme que, si *XCON* avait à l'origine remplacé soixante-quinze personnes au département chargé de la configuration, il en fallait à présent cent cinquante pour le mettre à niveau et le faire tourner (Crevier, 1997, p. 241).

Verbatim 1 : Historiquement, comment ils font pour faire de la publicité ? […] Avant, ils faisaient fonctionner leurs préjugés, c'est-à-dire qu'ils disaient qu'il y a un concept qui s'appelle la ménagère de moins de 50 ans et elle va être intéressée par le robot ménager *Turbo 2000*. Sauf qu'en réalité, […] il y a à peu près entre 35 et 40 % d'hommes. […] Quand on a des algorithmes qui sortent des résultats comme ça, on se demande si c'est l'algorithme qui déconne ou si c'est moi qui déconne ! Avec ce genre de méthode, vous allez apprendre de vos données. […] C'est phénoménal du point de vue de l'analyse des données. […] Vous allez faire émerger des réalités sociologiques à partir des données en faisant un problème d'optimisation assez simple. En jouant sur le *dataset* vous allez pouvoir apprendre des choses et vous allez ensuite pouvoir les réinjecter dans les modèles de données. […] Pour chaque goût vous avez un modèle. […] Vous regardez la variété du modèle et à quel moment votre performance plafonne. Il faut à peu près 500 paramètres en marketing (E_10).

Figure 2 : Extraits de matériau d'enquête

Le deuxième avantage est dans la continuité du premier. Il renvoie à l'idée que les technologies d'apprentissage artificiel seraient en capacité d'apprendre des connaissances immanentes au monde, c'est-à-dire qui, à la différence des humains (*cf.* Tversky & Kahneman, 1974), ne feraient l'objet d'aucun biais de représentation (*cf.* verbatim 1). Nous savons aujourd'hui que cette idée relève plus de mythe que de la réalité ; et cela, pour des raisons qui, en référence aux théories de la socialisation que propose la sociologie traditionnelle (et plus particulièrement celle de Pierre Bourdieu ; 1994), sont purement logiques. Car les réseaux de neurones artificiels sont conçus selon un modèle d'apprentissage connexionniste qui, sur le plan théorique, correspond à cette modélisation. À l'instar de la philosophie de l'esprit de John R. Searle (2004), il existe, par exemple chez Pierre Bourdieu (1994), l'idée que les institutions sociales sont incorporées, par le biais de la socialisation et sous la forme de structures mentales, dans les cerveaux des humains. Selon Jean-Pierre Changeux (2006), cette idée peut faire l'objet d'une explication d'ordre neurologique : c'est ce qu'il appelle la théorie des « bases neuronales de l'habitus ». Pour cet auteur, la socialisation aurait alors une forme matérielle dans la mesure où elle se manifesterait concrètement par à une grande diversité de configurations d'activation et d'inhibition de neurones biologiques. Or, si la théorie des bases neuronales de l'habitus est discutable[4], il n'en reste pas moins que les réseaux de neurones artificiels appliqués à la résolution de problématiques humaines, comme par exemple celles que pose la gestion de la relation client, la réalisent avec exactitude, à tout le moins d'un point de vue formel.

Attention : ne nous méprenons pas. Nos propos ne consistent ni à anthropomorphiser les machines intelligentes[5], ni à laisser penser que Pierre Bourdieu (1994) aurait œuvré à leur développement (ou

. . . . . .

[4] De nombreux travaux en sciences sociales et en psychologie cognitive (*cf.* Simon, 1996) montrent en effet que la cognition humaine n'est pas seulement composée d'une dimension physique et matérielle, mais aussi d'une dimension symbolique et fonctionnelle.

[5] Car, un peu à la manière dont Bruno Bachimont souligne l'idée que, « *plutôt que de considérer la pensée comme une manipulation algorithmique aveugle de symboles, il vaut mieux, au nom du sens, considérer ce que de telles manipulations nous permettent de penser* » (2000, p. 317), nous pensons que, plutôt que de considérer que les machines réalisent des formes d'apprentissages analogues à ceux des humains, il vaut mieux considérer qu'elles encapsulent des représentations humaines de ce qu'est l'apprentissage qui leur permettent de faire exister une forme de socialisation bien particulière qui, si

encore pour celui des philosophies de l'esprit qui en sont sous-jacentes). Il s'agit plus simplement de souligner la proximité qui existe entre le modèle connexionniste de l'apprentissage qu'encapsulent les réseaux de neurones et le modèle de socialisation que propose Pierre Bourdieu (1994) pour comprendre les mécanismes de reproduction social. Et si nous insistons sur cette proximité, c'est parce qu'elle permet de saisir une des plus graves problématiques éthiques associées au développement des technologies d'apprentissage artificiel au sein des organisations. La congruence qui existe entre le modèle connexionniste de l'apprentissage et le modèle de socialisation que propose Pierre Bourdieu (1994) est importante dans le sens où elle permet de comprendre, selon une perspective qui n'est pas d'ordre strictement technologique mais aussi sociologique, comment les systèmes d'apprentissage artificiel peuvent participer à la diffusion, par exemple, de stéréotypes sexistes ou racistes (*cf.* Caliskan, Bryson, & Narayanan, 2017) : les réseaux de neurone artificiels sont conçus d'une telle manière qu'ils incorporent au sein de leur architecture cognitive les biais de représentation que véhiculent les couples de données (x_i, y_i) qui les nourrissent et qui, avec l'avènement des *big data*, représentent des pans toujours plus grands de la société.

Aussi, bien que les réseaux de neurones ne soient pas les seuls systèmes d'apprentissage artificiel mobilisés par les acteurs socioéconomiques, un grand nombre de ces technologies recouvre, sur le plan purement formel, des mécanismes de production de connaissance dont les implications sociocognitives sont semblables. Le fonctionnement théorique des réseaux de neurones permet par là même de mieux comprendre comment, en autorisant l'automatisation de la communication organisationnelle, les formes de socialisation algorithmique que les technologies d'apprentissage artificiel autorisent peuvent contribuer à la reproduction de la société.

La socialisation algorithmique et la production de la société

Le fonctionnement des technologies d'apprentissage artificiel ne peut toutefois pas être compris par le seul biais de sa dimension théorique. Il renvoie également à une dimension empirique dans la mesure où ces technologies ont pour vocation d'être implémentées, à travers diverses applications, au sein de systèmes sociotechniques concrets. Cette dimension empirique est importante pour bien saisir les formes de socialisation algorithmique que permettent les technologies d'apprentissage artificiel.

Ainsi, l'enquête que nous avons menée sur les activités de conception de ces systèmes montre qu'elles constituent un processus de découverte d'un nouveau marché entre ceux qui les conçoivent et ceux qui les implémentent au sein de leurs organisations. Comme le montrent Sylvain Parasie et Éric Dagiral (2017), ce processus consiste alors, pour le concepteur, à aider l'implémenteur à découvrir et formaliser les connaissances dites métier qui sont nécessaires à la structuration du problème d'apprentissage que la machine doit résoudre. Ce travail d'exploration et de formalisation est opéré durant la conception de trois cadres sociocognitifs qui sont : l'environnement d'apprentissage, l'environnement de traitement et l'environnement politique. Comme nous l'avons souligné plus haut, l'objectif de cette section est de montrer comment, à travers la conception de ces trois cadres, les humains prêtent aux machines apprenantes divers opérateurs de traduction leur permettant de « re-présenter » le monde (Latour, 1996) et d'automatiser la communication selon un angle de vue qui apparaît pertinent aux premiers. Nous soutiendrons de ce fait que c'est par le biais

[suite de la note]

elle n'a empiriquement rien à voir avec celle des humains, n'en reste pas moins cruciale pour comprendre les phénomènes de (re)production sociale. Et c'est précisément en ce sens que nous parlons de socialisation algorithmique.

de ces trois cadres que les machines apprenantes participent à la création des entités cognitives qui forment la culture d'une société et, partant, encouragent plutôt sa production que sa reproduction.

La conception de l'environnement d'apprentissage

L'environnement d'apprentissage compose l'assemblage sociotechnique qui autorise la production et le recueil organisés de l'ensemble des données auxquelles la machine peut accéder. En référence aux travaux de Yves Jeanneret (2011 ; 2014), de Béatrice Galinon-Mélénec (2011) et à ceux de Jérôme Denis et Samuel Goëta (2017), cet assemblage est constitué de nombreuses couches sémio et socio-matérielles. Il forme un large dispositif de mise en trace des interactions que les hommes entretiennent avec les technologies numériques et qui s'étend jusqu'aux algorithmes de structuration des bases de données auxquelles les machines apprenantes sont connectées.

Comme le montrent les *verbatims* de la figure 3, la compréhension de cet assemblage par le concepteur est aussi difficile que fondamentale. Cette difficulté est notamment due au fait que : d'une part, les *data* des organisations constituent souvent un *patchwork* de données qui sont produites selon des finalités propres et qui peuvent être éloignées des buts associés aux développements des technologies d'apprentissage artificiel ; et, d'autre part, les personnes qui connaissent l'histoire de ces données sont souvent peu nombreuses et difficiles à identifier (*cf. verbatim* 2). Pour autant, comme l'indique le *verbatim* 3, la compréhension de cette histoire est cruciale parce qu'elle permet de mieux saisir les réalités que représentent les données, et, par voie de conséquence, les environnements à partir desquels les machines vont pouvoir effectuer leurs apprentissages. C'est d'ailleurs cette importance et cette complexité qui expliquent pourquoi certains des professionnels que nous avons interrogés préfèrent eux-mêmes poser les « sondes » qui autorisent la production et le recueil automatique des *big data*.

Verbatim 2 : Les organisations, elles ont leur histoire. Elles ont leur passé. Et cette histoire n'est pas *datacentric*, c'est-à-dire qu'elle n'est pas organisée autour de la donnée. Donc, les bases de données qui existent dans les organisations ne sont pas conçues pour faire de l'analyse de données. Elles sont conçues pour faire de l'administratif, pour faire de la facturation, pour faire des choses comme ça. Et donc, c'est un *patchwork* de dépôts de données. [...] En général, il y a des strates successives dans la donnée qui font qu'à l'instant *t*, c'est une accumulation de choses assez hétérogènes. Par exemple, une base de données dans laquelle on aurait ajouté des champs au cours du temps. Il y a des champs dans la base de données qui vivent, qui naissent, qui meurent, qui sont utilisés, puis plus utilisés. [...] Retrouver la personne dans l'organisation qui a l'histoire de la donnée, eh bien, c'est pas forcément évident. Peut-être que c'est une ou deux personnes seulement qui connaissent ça (E_01).

Verbatim 3 : Arriver dans l'entreprise, savoir quel type de données je dois aller collecter, quel service je dois aller voir, quel expert métier je dois aller voir, etc. Que je sois capable de juger de la qualité des données et de la représentativité des données. Savoir comprendre pourquoi j'ai collecté ses données. Qu'est-ce que ça veut dire que ces données soient là ? Et qu'est-ce que ça veut dire quand elles ne sont pas là ? Comprendre pourquoi il y a des non réponses. Qu'est-ce que ça veut dire ? Qu'est-ce qui a motivé, dans le comportement du consommateur, une non réponse ? Il y a une multitude d'informations qui ne sont pas uniquement mathématiques. [...] Si je me base [par exemple] sur mon auto-partage et que je collecte les données au niveau des bornes. Eh bien, je vais avoir 10 fois plus de gros clients que de petits clients. Parce que, les clients qui viennent 10 fois plus, par définition, je les vois 10 fois plus souvent. Alors que, si je me mets au niveau du fichier client, je vais avoir tout le monde. Je vais avoir autant les petits clients que les gros clients. Vous voyez, il y a plein de choses qui font que connaître les données et l'entreprise, ça parait primordiale pour mettre en place des modèles performants qui pourront être utiles demain (E_05).

Figure 3 : Extraits de matériau d'enquête

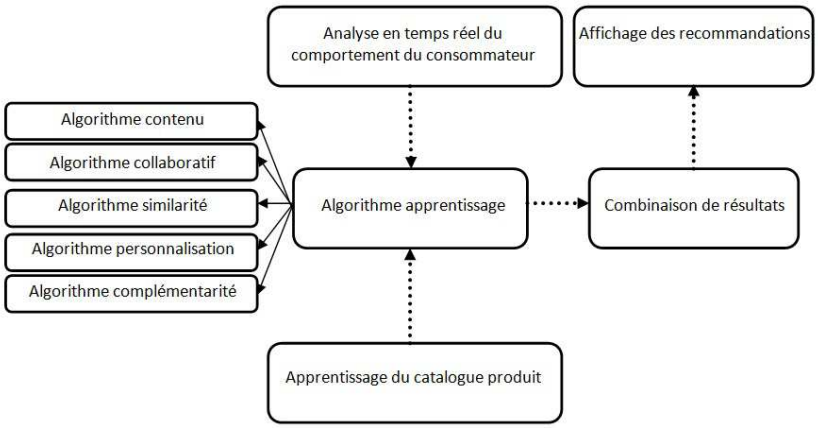
Les entretiens et les observations que nous avons effectués indiquent alors qu'une bonne partie des activités de conception des technologies d'apprentissage artificiel consiste à trouver et à développer le « bon » système de structuration automatique des données. Pour réaliser cette tâche, les concepteurs passent généralement un certain temps à visualiser les données, à échanger avec les experts métiers sur ce qu'ils peuvent comprendre de ces visualisations et à élaborer les techniques de traitement

automatique qui permettront de nettoyer les données, de remplir les valeurs manquantes, de recoder les variables et de les combiner entre elles. Dans leur ensemble, ces activités composent ce que les professionnelles nomment le « *feature engeneering* » (i.e., l'ingénierie des variables) et ont pour finalité d'organiser l'environnement d'apprentissage de l'algorithme en fonction de la façon dont son concepteur et son implémenteur comprennent la réalité que représentent les données ainsi que le rôle qu'ils souhaitent que la machine remplisse au sein de l'organisation. Pour l'ensemble des professionnels interviewés, la fabrication de l'environ d'apprentissage est donc centrale dans le sens où elle consiste à agencer la manière dont la machine conçoit le monde : ils savent bien que la pertinence des inférences que cette dernière pourra effectuer dépend étroitement de cette conception.

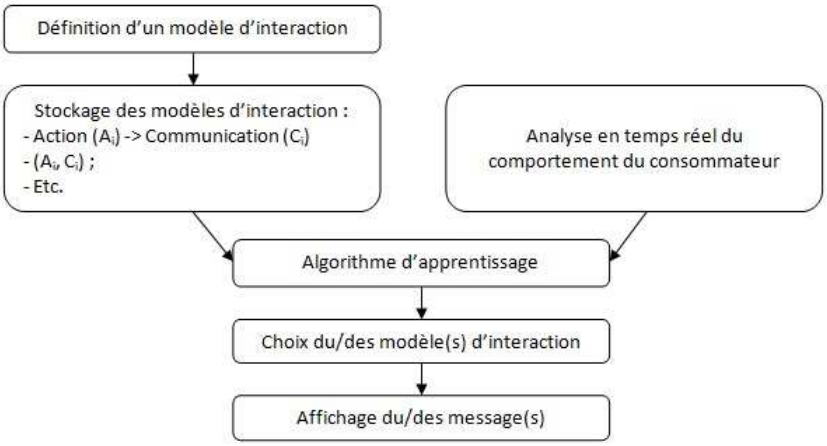
La conception de l'environnement de traitement

L'environnement de traitement constitue, quant à lui, l'architecture cognitive de la machine proprement dite. Sa conception renvoie à un travail qui est plutôt technique et théorique. En référence aux travaux de Herbert A. Simon (1996), ce travail consiste à élaborer le *design* du procédé qui permettra à la machine d'inférer des connaissances à partir des données dont elle dispose, c'est-à-dire à régler, à adapter, à fabriquer et/ou à combiner les algorithmes mobilisés (e.g., réseau de neurones, évolution simulée, ou encore, moteur d'inférence). Le rôle de l'implémenteur dans la conception de l'environnement de traitement est généralement de deux ordres. D'une part, l'implémenteur est souvent sollicité par le concepteur pour estimer l'utilité ou non de comprendre les apprentissages effectués par la machine. Si cela n'est pas nécessaire, le concepteur peut concevoir l'environnement de traitement à l'aide d'une ou plusieurs techniques d'apprentissage fonctionnant comme des boîtes noires[6]. D'autre part, si le concepteur pense utile de développer une architecture cognitive équipée d'un moteur d'inférence, l'implémenteur est mobilisé pour déterminer et formaliser les règles de déduction que la machine doit pouvoir mettre en action.

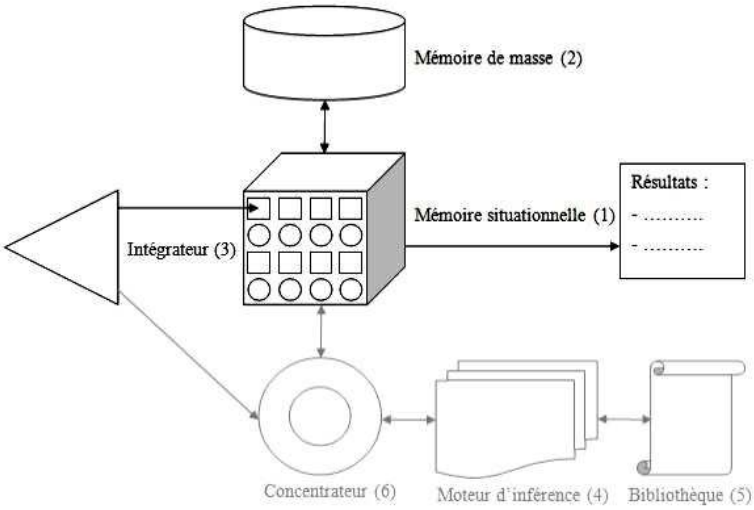
[6] C'est par exemple le cas des réseaux de neurones artificiels, et plus particulièrement des réseaux de neurones dits profonds ou convolutionnels, mais aussi des différentes techniques d'apprentissage par combinaison d'experts que sont les forêts aléatoires, le *boosting*, ou encore, le *bagging* (Cornuéjols & Miclet, 2010).



Architecture cognitive développée par la société S_01



Architecture cognitive développée par la société S_02



Architecture cognitive développée par la société S_03

Figure 4 : Exemples d'architectures cognitives appliquées à la gestion de la relation client

Comme l'indique la figure 4, il existe une grande diversité d'architectures cognitives. Celle élaborée par la société S_01 est par exemple composée de cinq algorithmes de recommandation dont la

sélection et la combinaison est supervisée par un algorithme d'apprentissage par renforcement. Cet algorithme d'apprentissage permet de décider quel est l'algorithme ou la combinaison d'algorithmes qui est la mieux adaptée en fonction du type de visiteur qui est en train de consulter le site, des évolutions des produits disponibles dans le catalogue de l'e-commerçant et du positionnement des espaces de recommandation à l'intérieur du site.

L'architecture cognitive conçue par la société S_02 est, quant à elle, constituée d'un moteur d'inférence et d'un système d'apprentissage par renforcement. Le moteur d'inférence regroupe plusieurs dizaines de modèles d'interaction qui sont définis avec les implémenteurs de cette technologie. Un de ces modèles est par exemple le suivant : si un consommateur clique *x* fois sur le même produit au cours des diverses visites du site qu'il a effectuées sur une période de *y* semaines, alors la machine doit activer la fenêtre *pop-up* lui présentant un code de promotion. L'algorithme d'apprentissage par renforcement défini alors deux types de modalité d'activation des règles d'inférence. D'une part, il détermine les modalités d'application des modèles d'interaction du moteur d'inférence. Par exemple, dans le cas de la règle que nous venons de mentionner, cet algorithme défini les valeurs de *x* et de *y* qui sont économiquement les plus performantes par une série d'essais-erreurs. D'autre part, il cherche à identifier, toujours par le biais de multiples essais-erreurs, à quel point chaque règle stockée dans le moteur d'inférence est plus ou moins efficace en fonction des profils des consommateurs.

Comme nous l'avons mentionné plus haut, la société S_03 a développé une architecture cognitive différente de celles que nous venons de présenter dans la mesure où celle-ci est dotée d'un module d'apprentissage qui s'apparente à un réseau de neurones profond et d'un module d'inférence. Le fonctionnement de ce système est complexe (*cf.* Vayre, 2018). Dans les grandes lignes, le module d'apprentissage permet d'induire des situations du type : « *les consommateurs qui ont acheté le produit x ont également acheté le produit y* ». Ces situations sont ensuite stockées sous forme de règles dans le moteur d'inférence qui peut de ce fait décider de déclencher la recommandation du produit *y* à chaque fois qu'un consommateur achète le produit *x*, et inversement. Mais le moteur d'inférence est également souvent utilisé par l'implémenteur pour contrôler l'architecture cognitive de la machine en plaçant en son sein les situations d'activation des stratégies de communication qu'il souhaite que la machine réalise.

La conception de l'environnement politique

L'environnement politique est composé de l'ensemble des critères qui permettent à la machine de contrôler la performance de ses apprentissages. Aussi, même s'il existe des systèmes d'apprentissage qui n'ont pas besoin de ces critères pour fonctionner (*cf.* les technologies dites non-supervisées ; Cornuéjols & Miclet, 2010), ceux-ci sont généralement associés, au moins en bout de chaîne, à d'autres techniques qui les nécessitent. Les machines ne peuvent en effet souvent rien apprendre d'utile aux yeux de ceux qui les implémentent sans que leurs concepteurs incorporent au sein de leur architecture cognitive les critères qui composent ce que nous appelons l'environnement politique. Et si nous proposons une telle dénomination, c'est parce que le système d'évaluation de la performance que cet environnement constitue a pour principale fonction de donner un sens à l'autonomie cognitive de la machine dans la mesure où il oriente la totalité du travail inférentiel qu'elle effectue.

Par exemple, dans le cas de la société S_03, ces critères de performance sont les indices de préférence que nous avons exposés plus haut, c'est-à-dire les notations que les consommateurs confèrent aux produits du e-commerçant, l'attention qu'ils leur accordent en termes de durée de consultation et l'intérêt qu'ils leur portent en les plaçant ou non dans leurs paniers. Il est alors important de bien comprendre qu'au-delà du fait que la valeur indiciaire de ces critères de préférence soit discutable, les représentations que la machine apprenante de S_03 se construit du monde seraient complètement différentes si son environnement politique ne consistait plus à

maximiser les préférences des consommateurs, mais à minimiser leur désorientation. Au regard des entretiens et des observations que nous avons pu effectuer, l'environnement politique est, à l'inverse de l'environnement de traitement, principalement défini par l'implémenteur.

Verbatim 4 : Nous, on peut jouer avec les données, on peut trouver beaucoup de choses. Mais, s'il n'y a pas un sens derrière pour celui qui les utilise… Donc, c'est des jeux de va-et-vient. On travaille sur nos données, on en tire des paramètres, des explications, des moyennes plus ou moins complexes des données. Puis après, on retourne vers le client qui reconnaît ou qui découvre des choses propres à son métier. Et à partir de là, on peut ré-avancer : soit aller chercher plus loin, soit faire de la prédiction […]. Et donc, entre notre expertise statistique et le client, on redéfinit qu'elle est le problème. Qu'est-ce que c'est qu'une moyenne ? Comment on structure les données ? Essayer de réduire les dimensions, on se met en groupe homogène, on travaille sur des sous-ensembles, et après, on teste des méthodes plus ou moins évoluées de prédiction en fonction des besoins du client. [… C'est comme ça qu'on réalise] le travail de structuration [qui] consiste à trouver par quel biais on veut regarder le problème. C'est-à-dire que, si on a quelque chose qui va varier en fonction du temps, en fonction du lieu, en fonction du sexe de la personne et en fonction du temps qu'il fait, quand on cherche à savoir comment est-ce qu'il varie, on va regarder plusieurs observations et on va regarder leur différence. Mais, est-ce que je regarde la différence au même endroit pour une femme dans un environnement ensoleillé et je regarde juste deux jours ; ou alors, le même jour à deux endroits différents pour une femme et un homme ? Alors, ça fait déjà trop de différence, parce qu'est-ce que la différence est due au fait que c'est une femme, au fait qu'il fait moche, etc. ? Donc, on commence par structurer les données en fonction du problème qu'on peut extraire. Et ça, ce n'est pas quelque chose qui est simple, qui est automatique, ça demande pas mal d'échanges avec les experts métier (E_02).

Figure 5 : Extraits de matériau d'enquête

De manière plus générale et en référence au *verbatim* 4 exposé ci-dessus, la conception d'une technologie d'apprentissage artificiel recouvre un travail de structuration qui consiste à trouver le biais à partir duquel une machine peut se représenter la réalité. Ce biais d'apprentissage résulte de la composition des différents choix qui sont co-élaborés par le concepteur et l'implémenteur pour développer les environnements d'apprentissage, de traitement et politique que nous venons de présenter. Du point de vue du concepteur et de l'implémenteur, c'est ce biais qui est garant de la pertinence des apprentissages que la machine doit réaliser : en orientant les formes de sa socialisation algorithmique, il assure l'utilité socioéconomique de son travail inférentiel et, par voie de conséquence, celle des informations qu'elle communique au sein de l'organisation. C'est précisément en ce sens qu'il faut comprendre qu'en automatisant la communication organisationnelle et sur le plan purement empirique, les technologies d'apprentissage artificiel produisent plus qu'elles ne reproduisent les institutions sociales que représentent les données qui les nourrissent, qu'elles incorporent au sein même des règles d'inférence qu'elles fabriquent et qui composent leur architecture cognitive.

Conclusion

En automatisant la communication au sein des organisations, comment les machines apprenantes peuvent-elles participer à la (re)production de la société ? À partir du cas des réseaux de neurones artificiels, nous avons vu que, sur le plan théorique, les technologies d'apprentissage artificiel appliquées à la gestion de la relation client autorisent l'automatisation de la communication organisationnelle selon des mécanismes de socialisation algorithmique qui sont plutôt favorables à la reproduction de la société. D'un point de vue purement formel, ces technologies ont pour objet de traduire les structures sociales ou, si l'on préfère, les régularités comportementales que représentent les *big data* en des structures cognitives qui, par exemple dans le cas des réseaux de neurones artificiels, prennent la forme concrète de configurations d'activation et d'inhibition des unités de calcul qui les composent. Pour autant, nous avons vu que, sur le plan empirique, les technologies d'apprentissage artificiel recouvrent un travail de conception de trois cadres sociocognitifs qui orientent les formes de la socialisation algorithmique des machines. Dans leur ensemble, ces trois cadres constituent le biais qui permet aux humains de contrôler les formes des apprentissages de la machine et, partant, celles des informations qu'elle fabrique et communique aux entités sociales et

techniques qui composent l'organisation. Autrement dit, ce biais d'apprentissage prend la forme concrète d'un ensemble d'opérateurs de traduction qui transforment les régularités que représentent les *big data* en des informations qui doivent autoriser l'automatisation de la communication organisationnelle. Tout l'enjeu de la conception d'une machine apprenante est alors que cette automatisation apparaisse pertinente du point de vue du concepteur, et surtout, de celui de l'implémenteur. C'est donc en ce sens qu'en référence aux travaux de Yves Jeanneret (2008), les formes de communication automatique qu'instituent les machines apprenantes comportent une dimension logistique, social, mais aussi, poétique : les informations que ces machines produisent et communiquent sont des créations qu'elles réalisent à partir des masses de données qui les nourrissent et en fonction de la façon dont ceux qui les conçoivent et qui les implémentent organisent leurs environnements d'apprentissage, de traitement et politique.

Références bibliographiques

Akrich, Madeleine (1987), « Comment décrire les objets techniques ? », Techniques et culture, n° 9, p. 49-64.

Akrich, Madeleine (1989), « La construction d'un système socio-technique : esquisse pour une anthropologie », Anthropologie et sociétés, vol. 13, n° 2, p. 31-54.

Akrich, Madeleine (2004), « Comment décrire l'interaction entre les techniques et les humains ? » (p. 159-178), in Akrich, Madeleine ; Callon, Michel ; Latour Bruno, Sociologie de la traduction : textes fondamentaux. Paris : Presses des Mines.

Bachimont, Bruno (2000), « L'intelligence artificielle comme écriture dynamique : de la raison graphique à la raison computationnelle » (p. 290-319), in Petitot, Jean ; Fabbri Paolo, Au nom du sens. Autour de l'œuvre d'Umberto Eco, Paris : Grasset.

Bandura, Albert (1986), Social Foundationd of though and action: a social cognitive, Englewood Cliffs: Prentice-Hall.

Bouillon Jean-Luc (2015), "Technologies numériques d'information et de communication et rationalisations organisationnelles : les « compétences numériques » face à la modélisation", *Les Enjeux de l'Information et de la Communication*, n°16/1, p. 89 à 103, [en ligne] https://lesenjeux.univ-grenoble-alpes.fr/2015/06-Bouillon/

Boullier, Dominique (2016), Sociologie du numérique, Paris : Armand Collin.

Bourdieu, Pierre (1970), La reproduction : éléments d'une théorie du système d'enseignement, Paris : Éditions de minuit.

Bourdieu, Pierre (1994), Raisons pratiques. Sur la théorie de l'action, Paris : Seuil.

Caliskan, Aylin ; Bryson, Joanna J. ; Narayanan, Arvind (2017), « Semantics derived automatically from language corpora contain human-like biases », Science, vol. 356, n° 6334, p. 183-186.

Callon, Michel (1998), The laws of the markets, Oxford: Blackwell.

Cardon, Dominique (2015), A quoi rêvent les algorithmes. Nos vies à l'heure des big data, Paris : Seuil.

Cardon, Dominique ; Casilli, Antonio (2015), Qu'est-ce que le Digital Labor ?, Paris : Édition de l'INA.

Changeux, Jean-Pierre (2006), « Les bases neurales de l'habitus » (p. 143-158), in Fussman, Gérard (dir.), Croyance, raison et déraison, Paris : Odile Jacob.

Cooren, François ; Fairhurst, Gail T. (2009), « Dislocation and stabilization: how to scale up from interactions to organizations » (p. 117-152), in Putnam, Linda L. ; Nicotera, Anne Maydan (dir.), Building theories of organizations: the constitutive role of communication, New York: Routledge.

Cornuéjols, Antoine ; Miclet, Laurent (2010), Apprentissage artificiel : concepts et algorithmes. Paris : Eyrolles.

Crevier, Daniel (1997), À la recherche de l'intelligence artificielle, Paris : Flammarion.

Dagiral, Éric ; Parasie, Sylvain (2017), « La « science des données » à la conquête des mondes sociaux : ce que le « Big Data » doit aux épistémologies locales » (p. 85-104), in Menger Pierre-Michel ; Paye, Simon (dir.), Big data et traçabilité numérique : les sciences sociales face à la quantification massive des individus, Paris : Collège de France.

Denis, Jérôme ; Goëta, Samuel (2017), « La fabrique des données brutes. Le travail en coulisses de l'open data » (p. [en ligne]), in Mabi, Clément ; Plantin, Jean-Christophe ; Monnoyer-Smith, Laurence, Ouvrir,partager, réutiliser. Regards critiques sur les données numériques, Paris : Éditions de la Maison des sciences de l'homme.

Denis, Jérôme ; Pontille, David (2010), « Performativité de l'écrit et travail de maintenance », Réseaux, vol. 5, n° 163, p. 105-130.

Dujarier, Marie-Anne (2008), Le travail du consommateur. De McDo à eBay : comment nous coproduisons ce que nous achetons, Paris : La Découverte.

Forgy, Charles L. (1981), OPS5 user's manual. tech. rep., Pittsburgh: Computer science department of the Carnegie Mellon University.

Galinon-Mélénec, Béatrice (2011), L'Homme trace. Perspectives anthropologiques des traces contemporaines, Paris : CNRS.

Ganascia, Jean-Gabriel (2017), Intelligence artificielle : vers une domination programmée ?, Paris : Le cavalier bleu.

Giddens, Anthony (1986), The Constitution of Society: outline of the theory of structuration, California: Berkeley.

Hebb, Donald O. (1949), The organization of behavior, New York: Wiley.

Jeanneret, Yves (2008), Penser la trivialité. Volume 1 : la vie triviale des êtres culturels, Paris : Hermès-Lavoisier.

Jeanneret, Yves (2011), « Les harmoniques du Web : espaces d'inscription et mémoire des pratiques », Médiation et Information, n° 32, p. 31-40.

Jeanneret, Yves (2014), « La fabrique de la trace : une entreprise herméneutique » (p. 47-64), in Idjéraoui-Ravez, Linda ; Pélissier, Nicolas (dir.), Quand les traces communiquent... Culture, patrimoine, médiatisation de la mémoire, Paris : Harmattan.

Jézégou, Annie (2014), « L'agentivité humaine : un moteur essentiel pour l'élaboration d'un environnement personnel d'apprentissage », STICEF, n° 21, p. 269-286.

Kohler, Dorothée ; Weisz, Jean-Daniel (2016), « Industrie 4.0 : comment caractériser cette quatrième révolution industrielle et ses enjeux ? », Annales des Mines - Réalités industrielles, n° 4, p. 51-56.

Latour, Bruno (1994), « Une sociologie sans objet ? Remarques sur l'interobjectivité », Sociologie du travail, vol. 36, n° 4, p. 587-607.

Latour, Bruno (1996), « Le « pédofil » de Boa Vista – montage photo-philosophique » (p. 171-225), in Latour, Bruno, Petites leçons de sociologie des sciences, Paris : Seuil.

Laumond, Jean-Paul (2016), « La robotique », Annales des Mines - Réalités industrielles, n° 4, p. 43-46.

Maistre, Christophe (de) (2016), « L'usine cyberphysique : usine connectée, simulée et reconfigurable », Annales des Mines - Réalités industrielles, vol. 4, p. 37-42.

McCulloch, Warren S. ; Pitts, Walter (1943), « A logical calculus of the ideas immanent in nervous activity », The bulletin of mathematical biophysics, vol. 5, n° 4, p. 115-133.

Mead, George H. (2015), Mind, self, and society, Chicago: University of Chicago Press.

Miège, Bernard (2007), « Sur le positionnement de la recherche en histoire des SIC », Questions de communication, n° 12, p. 191-202.

Musso, Pierre (2017), La religion industrielle. Monastère, manufacture, usine. Une généalogie de l'entreprise, Paris : Fayard.

Parsons, Talcott E. (2005), The Social System, New York: Routledge.

Pickering, Andrew (1995), The mangle of practice: time, agency, and science, Chicago: University of Chicago Press.

Rosenblatt, Frank (1958), « The perceptron: a probabilistic model for information storage and organization in the brain », Psychological Review, vol. 65, n° 6, p. 386-408.

Searle, John R. (2004), Mind: a brief introduction, Oxford: Oxford University Press.

Simon, Herbert A. (1996), The sciences of the artificial, Cambridge: MIT Press.

Tiffon, Guillaume (2013), La mise au travail des clients, Paris : Economica.

Tversky, Amos ; Kahneman, Daniel (1974), « Judgment under uncertainty: heuristics and biases », Science, vol. 185, n° 4157, p. 1124-1131.

Vayre, J.-S. (2016), Des machines à produire des futurs économiques : sociologie des intelligences artificielles marchandes à l'ère du big data, Toulouse : Université Toulouse Jean Jaurès.

Vayre, J.-S. (2018), « Machines intelligentes et économie numérique : étude du cas d'un agent artificiel dans le domaine du travail relationnel marchand », Les Cahiers du Numérique, vol. 14, n° 1, p. 83-109.

Les données de la guerre. Big Data et algorithmes à usage militaire

Data of the war. Big Data and algorithms for military use

Los datos de la guerra. Big Data y algoritmos para uso militar

Article inédit, mis en ligne le 15 novembre 2018.

Olivier Koch

Olivier Koch est enseignant à l'Université Galatasaray dans le département de communication. Il est rattaché au laboratoire des Sciences de l'information et de la communication (LabSIC) de l'Université Paris 13. Ses travaux portent sur les enjeux de l'information dans des contextes de guerre et sur la réforme des médias en contexte de recomposition politique.

Plan de l'article

Résumé

Les dispositifs « *Big Data* et algorithmes » ont été intégrés au secteur militaire américain dans les années 2000. En Afghanistan et en Irak, les armées les ont utilisés pour détecter les « insurgés » au sein de population et pour prédire de nouvelles « insurrections ». Ces dispositifs ont été employés dans le but d'automatiser la détection et la prédiction, et optimiser ainsi la prise de décision politique. Cependant, malgré les efforts institutionnels et financiers consentis, la promesse de doter le chef de guerre d'une ingénierie plus efficiente que les précédentes n'a pas été tenue. Et cet échec n'a pas remis en question la perpétuation de ces programmes. Nous proposons dans cet article d'analyser l'intégration des dispositifs « *Big Data* et algorithmes » au secteur militaire étasunien au prisme cette contradiction.

Mots clés

Big Data, contre-insurrection, comportements socioculturels, populations, prise de décision.

Abstract

The "Big Data and Algorithms" devices were integrated into the US military in the 2000s. In Afghanistan and Iraq, the armed forces used them to detect "insurgents" within the population and to

predict new "insurrections". These devices have been used to automate detection and prediction, thereby optimizing political decision-making. However, despite the institutional and financial efforts made, the promise to equip the warlord with more efficient engineering than the previous ones was not fulfilled. And this failure did not call into question the perpetuation of these programs. In this article, we propose to analyze the integration of "Big Data and Algorithms" devices with the US military sector through this contradiction.

Keywords

Big data, counterinsurgency, culture, sociocultural behavior, decision-making

Resumen

Los dispositivos de "Big Data y algoritmos" se integraron en el sector militar americano en la década de 2000. En Afganistán y en Iraq, las fuerzas armadas los usaron para detectar "insurgentes" dentro de la población y predecir nuevas "insurrecciones". Estos dispositivos se han utilizado para automatizar la detección y predicción, optimizando así la toma de decisiones políticas. Sin embargo, a pesar de los esfuerzos institucionales y financieros realizados, la promesa de equipar al caudillo con una ingeniería más eficiente que las anteriores no se cumplió. Y este fracaso no cuestionó la perpetuación de estos programas. En este artículo, proponemos analizar la integración de los dispositivos de "Big Data and algoritmos" en el sector militar estadounidense a través de esta contradicción.

Palabras clave

Big Data, contrainsurgencia, cultura, comportamiento sociocultural, toma de decisiones

Introduction

« *La culture est le "terrain humain" de la guerre, le terrain humain est le terrain clé* [1] ». Cette déclaration du Major General Geoffroy Lambert restitue l'orthodoxie qui a guidé les réformes du secteur militaire étatsunien dans les années 2000. Durant cette décennie, les armées en Irak et en Afghanistan ont renoué avec la contre-insurrection, un art de la guerre dans lequel la détection des « insurgés » parmi le reste de la population est au centre des enjeux tactiques. Afin de s'orienter sur ce « terrain humain », le département de la Défense a investi dans le traitement automatisé de *data* sur les comportements socioculturels des populations autochtones. Les dispositifs « *Big Data* et algorithmes » ont ainsi été utilisés pour mener une guerre « centrée sur la culture » (Scale, 2004).

L'intégration de ces dispositifs au sein du secteur militaire étasunien fait l'objet de cet article. On se propose d'analyser ses formes et les logiques qui l'ont gouvernée à travers deux niveaux de ses manifestations concrètes. Tout d'abord, les programmes mis en œuvre pour guider les armées sur les terrains de la guerre irrégulière. Il s'agit, à ce premier niveau, de signaler quelles nouvelles technologies ont été incorporées aux systèmes de détection et de prédiction automatisées, avec quels

. . . . . .

[1] Cette citation est tirée de l'article de Yuri Lecchuk et d'Alexander Lubyansky, « Cultural agent model to predict inhabitant opinion réactions (CAMPHOR) : building and applying a dynamic humain terrain map », issu de la conférence « 13eICCRTS C2 for complex behavior endeavor », [en ligne], Consulté le 2 février 2018, http://www.dodccrp.org/events/13th_iccrts_2008/CD/html/papers/156.pdf

effets de performance escomptés, puis de restituer leurs usages dans une économie du recours à la violence armée. A un deuxième niveau, l'intégration de ces dispositifs dans le secteur militaire est appréhendée à travers la réorganisation institutionnelle de la Recherche et Développement dont ont émergé les programmes. Cette réorganisation permet de saisir, en particulier, les dynamiques de convergences intersectorielles orchestrées au plus haut niveau de l'Etat fédéral.

La base documentaire exploitée dans cette investigation est composée de deux types de documents, différenciés par l'origine sectorielle de leurs auteurs. Une partie du corpus agrège des publications éditées par des organisations du département de la Défense. On a cherché à y relever les mobiles et les justifications du recours aux ingénieries des *data.* L'autre partie est composée de publications de chercheurs-ingénieurs en sciences computationnelles des données. Le travail de valorisation des performances de leurs recherches appliquées y a fait l'objet d'une attention particulière. En croisant les justifications des premiers et les valorisations des seconds, il s'agit de mettre en évidence la convergence de logiques d'acteurs — celles du politique (le chef des armées) et celles de l'ingénieur — sur un même horizon d'attente : l'optimisation de la prise de décision politique.

Cet article se compose de trois parties. La première porte sur la mise en œuvre de systèmes de détection du « terroriste » ou de l'« insurgé » au sein des populations afin de guider les drones vers leurs cibles. A l'aune de la réorganisation de la Recherche et Développement, la deuxième est consacrée à l'évolution des machines prédictives dédiées à l'anticipation des crises imminentes. Enfin, dans la dernière partie, on se propose de restituer les innovations technologiques de la guerre « centrée sur la culture » à partir d'une « économie de la promesse » (Joly, 2010), la promesse d'optimiser la prise de décision politique, puis de monter en quoi cette promesse n'est ni tenue, ni tenable.

Détecter et « neutraliser »

Depuis l'avènement de la cybernétique, l'informatique connectée a été employée pour parfaire les systèmes de guidage des armes vers leurs cibles (Forget, 2001). Avec le tournant contre-insurrectionnel en Irak, les dispositifs « *Big Data* et algorithmes » ont été employés à la détection des cibles humaines au sein des populations autochtones. Dans cette première partie, il s'agit d'analyser les modalités de fonctionnement et d'application de ces outils de détection en restituant leur emploi dans une économie du recours à la violence armée.

Comprendre le recours aux technologies de détection automatisée implique de saisir les spécificités de la guerre contre-insurrectionnelle. Dans ce type de conflit, l'ennemi se confond avec le reste de la population : il n'est pas nécessairement armé et en tenue militaire, et il évolue dans les milieux urbains comme d'autres citadins. De ce point de vue, la détection de l'« insurgé » dans le brouillard de la guerre irrégulière a deux enjeux concomitants. Au niveau tactique, il importe de repérer cet ennemi pour neutraliser ses capacités de résistance. Au niveau stratégique, l'enjeu est de le distinguer du reste des civils pour éviter de tuer des individus sans lien avec l'ennemi. Le risque serait d'accroître l'hostilité des populations vis-à-vis de l'occupation et de contribuer ainsi à la légitimation des groupes armés « insurrectionnels ». Or, de cette légitimité dépend le soutien des populations à ces groupes. Les dispositifs « *Big Data* et algorithmes » ont été utilisés dans les années 2000 pour modéliser les comportements de cet « insurgé » (ou « terroriste ») et pour optimiser ainsi la prise de décision du commandement.

Afin de parfaire cette modélisation, le département de la Défense a réorganisé ses activités en Recherche et Développement. En 2009, a été créé le *Human Social Culture Behavior program* (HSCB)[2] au sein duquel a été a été mis en œuvre le *Cultural Knowledge Consortium* (CKC). La mission de ce consortium était de développer des recherches en modélisation algorithmique des

[2]Cf. Human socio cultural behaviour modeling Program, [En ligne], Consulté le 19 juillet 2017, https://info.publicintelligence.net/DoD-SocioculturalBehavior.pdf

comportements socioculturels. A partir de 2010, le CKC est devenu le *Global cultural knowledge network* avec pour objectif de porter à une autre échelle la production et la coordination de ces savoirs. Cette organisation dont la mission est « *de rassembler toute la capacité intellectuelle des États-Unis [...] en guidant la connaissance socioculturelle vers la décision*[3] » est conçue comme l'infrastructure d'une communauté de recherche transnationale. Le secteur académique national a été également mis à contribution. Les dynamiques au sein des Universités ont été insufflées en 2008 par le projet *Minerva Research Initiative*, grâce à des financements visant à orienter les travaux en sciences sociales sur la cartographie-analyse de réseaux humains. Avec ces technologies de profilage des populations, déjà analysées par Armand Mattelart et André Vitalis (2014), une « nouvelle physique sociale » (Pucheu, 2017) de la déviance et du crime a vu le jour.

L'un des programmes les plus représentatifs de « la guerre centrée sur la culture » dans la deuxième moitié des années 2000 est le « *Human Terrain* System » (HTS) (« système du terrain humain »)[4]. Il est défini par ses entrepreneurs comme l'ensemble des « *éléments sociaux, ethnographiques, culturels, économiques et politiques des populations à travers lesquelles une force opère* », autant d'éléments qu'il s'agit de transformer en données numériques pour qu'elles soient « utilisées dans une partie du processus de prise de décision militaire » (Kipp *et al.*, 2006, p. 9). Le HTS est employé à la détection automatisée d'« insurgés » au sein des populations autochtones. Cette détection est fondée sur l'analyse comportementale d'individus appréhendés dans leurs réseaux affinitaires. Qui fait quoi ? Où ? Quand ? Et avec qui ? Voici en somme les principales questions auxquelles le système doit figurer les réponses sous formes de graphes. Selon la position nodale d'individus dans des réseaux estimés hostiles, leur « neutralisation » peut être décidée par les chefs de guerre, soit par mise à mort, soit par mise en détention.

Les informations et les *data* primaires (les *inputs*) du système sont collectées par observation humaine à travers des missions de renseignement —par des soldats et des anthropologues « *embedded* »— et à travers la surveillance automatisée des télécommunications locales et au moyen de caméras fixées sur des drones. Renseignement humain et renseignement automatisé sont ainsi conjugués dans la constitution des gisements de *data* exploités par le HTS. Les données de ces gisements, systématiquement enregistrées dans un *data center* aux Etats-Unis, sont combinées dans le but d'identifier des groupes et des « *ego networks* » dont le logiciel de *Mapping Human Terrain* (MAP-HT) produit la représentation. Cette cartographie sociale (le *output*) doit permettre aux armées de s'orienter dans le brouillard de la guerre irrégulière. Elle est constituée de trois calques superposés. Le premier calque figure des groupes et réseaux primaires définis selon des attributs qualifiés de « culturels » (dans le lexique des artisans du HTS) : religion, tribu, ethnie, clans. Le deuxième calque représente sous forme de graphe le « réseaux intégré » des relations de parenté, des relations affinitaires et des communications routinières entre individus. Enfin, à partir d'une fusion des deux premiers, le troisième calque superpose les graphes obtenus à la géographie physique des lieux grâce à un logiciel de géolocalisation.

Le HTS schématise des « formes de vie » selon un procédé décrit par Grégoire Chamayou dans *Théorie du drone* (2016). Ces « formes » sont élaborées par l'analyse de données sur ce que font quotidiennement les individus dans leurs réseaux sociaux. De la récurrence de leurs activités émergent des *patterns* repérables (des « schémas de vie »). Au regard de ces récurrences, les comportements qui dévient de cette trame signalent une menace. Ces comportements déviants sont estimés en effet « signer » une appartenance à un réseau hostile, à un groupe d'« insurgés » ou une organisation « terroriste ». Dès lors, leur détection engage des « frappes par signature » via des drones, une mise à mort donc. Dans ce cas, la guerre cynégétique est avant tout une « chasse à l'homme » où, contrairement au modèle clausewitzien du duelliste, l'un des belligérants (le « terroriste ») ne se bat plus contre son adversaire : il devient une proie (*Ibid.*).

. . . . . .

[3] [En ligne]. Consulté le 25 juin 2017, https://community.apan.org/wg/oekn/

[4] Mis en œuvre entre 2005 et 2006, officiellement le HTS aurait pris fin en 2014. Cependant, selon Benjamin D. Hopkins (2016), l'armée américaine a été contrainte d'avouer publiquement que le projet a été repris et intégré dans un nouveau programme de recherche (le *Global Cultural Knowledge Network*).

Le *Humain terrain system* a été utilisé dans la contre-insurrection en Irak et en Afghanistan, mais cette technologie est aussi devenue dans les années 2010 une pièce importante de la « guerre à distance » (« *remote warfare* »). Ce modèle opérationnel a servi de référence aux armées américaines et britanniques en Afrique, dans la lutte contre Aqmi et Boko Haram, et au Moyen-Orient dans la guerre contre l'« État islamique ». Aux États-Unis, cette « guerre à distance » a singularisé les opérations militaires menées sous la présidence de Barack Obama (2009-2017), en rupture avec la logique d'occupation des territoires mise en œuvre en Irak sous le mandat de Georges W Bush (2001-2009). De ce point de vue, l'informationnalisation[5] de la guerre dans les années 2010 est au service d'une projection de la force sans déploiement d'hommes au sol. Des déploiements qui exposent les soldats au feu de l'ennemi et augmentent le nombre de « morts en opération », ce dont le gouvernant devrait rendre compte devant les publics domestiques. Les technologies de détection de la « guerre à distance » sont ainsi employées à la réduction des coûts politiques de la guerre.

Le dispositif contribue, sous un autre rapport, à l'économie du gouvernement des hommes dans les chaînes de commandement. L'automatisation de la détection transfère la responsabilité d'un interprète humain aux productions anonymes des graphologies numériques. Dans la concaténation des actes et des décisions qui mènent à la mise à mort, une partie de la responsabilité est ainsi déléguée aux machines détectives. Le commandement décide de l'exécution d'un homme, un soldat actionne le drone à distance derrière son écran mais, dans un cas comme dans l'autre, l'acteur est affranchi d'un travail d'interprétation des données (dont dépend l'imputation *a posteriori* des culpabilités liées aux crimes de guerre cynégétiques). Le dispositif permet de limiter ainsi les jugements que les différents opérateurs de la force armée pourraient prononcer sur l'usage de la violence « légitime » d'État. Or, ces jugements peuvent susciter chez ces opérateurs des hésitations, des refus d'obtempérer, et peuvent affecter ainsi la conduite de la guerre. La robotisation de la guerre sur les zones d'affrontement va dans le même sens. Elle pourvoit les armées de soldats-machines qui ne discuteront pas les ordres et qui ne refuseront jamais d'aller au combat (Singer, 2009). Sous ce rapport, l'intermittence homme-machine dans la division du travail militaire n'est pas seulement destinée à optimiser l'efficacité du déploiement de la force armée : elle agence le gouvernement des hommes.

Prédire et anticiper les « instabilités »

Conjointement au développement des machines détectives, le département de la Défense a financé la conception de programmes destinés à automatiser la prédiction de crises. Dans la lignée de systèmes mis en œuvre dans les décennies précédentes, l'objectif était de prédire des « instabilités » sociopolitiques et d'anticiper ainsi une reconfiguration des échiquiers géopolitiques. Dans cette deuxième partie, on se propose d'appréhender comment les machines prédictives ont évolué avec l'intégration des dispositifs « *Big Data* et algorithmes » dans le secteur militaire et, plus particulièrement, avec la redéfinition des enjeux stratégiques de la « guerre centrée sur la culture ».

Aux États-Unis, la « politique des oracles » est constituée en secteur professionnel depuis la guerre froide (Colonomos, 2014). Dans ce secteur, l'agence fédérale en charge du développement des nouvelles technologies à usage militaire (DARPA) a financé des systèmes de prédiction automatisée au milieu des années 1970, notamment dans le cadre du *Integrated Crisis Early Warning System* (ICEWS) (Andriole, Young, 1977). A suivre les analyses de Sean O'Brien (2002), à partir du milieu des années 1990 la prédiction automatisée s'est progressivement appuyée sur l'usage actuariel des sciences computationnelles et des algorithmes prédictifs. Conçu en 2005, le projet *Senturion* illustre cette évolution. Développé à l'Université de la défense nationale, il a été mis en œuvre dans le but de prédire les comportements de décideurs politiques (individus et groupes) dans des pays étrangers. Qui va faire quoi et avec quels effets ? Le programme répondait à ces questions en utilisant des

[5] L'« informationnalisation » de la guerre désigne le processus d'encodage d'existences et de pratiques qui entrent dans le périmètre de la rationalité stratégique.

« *data sur des échantillons de décideurs* », et « *des algorithmes représentant des processus comportementaux* » (Abdollahian, Baranick, Efird, Kuger, 2006).

Un tournant a été amorcé dans la deuxième moitié des années 2000. La prédiction automatisée a été progressivement réorganisée selon les enjeux stratégiques de la « guerre centrée sur la culture ». Deux inflexions majeures marquent cette réorganisation. Le *Integrated Crisis Early Warning System* (ICEWS) qui avait vu le jour pendant la guerre froide a été intégré en 2009 au *Human Social Culture Behavior program* (HSCB). L'insertion du ICEWS dans le HSCB acte ainsi la volonté de l'administration américaine de faire évoluer les machines prédictives en exploitant des *Big Data* « socioculturelles ». Une seconde inflexion dans l'évolution des dispositifs prédictifs, consécutive à la première, s'est traduite par l'exploitation de ce type de *data* à partir de gisements de données issues des médias et des réseaux sociaux numériques (RSN). On se propose dans ce qui suit de restituer cette évolution à travers l'automatisation du codage des données médiatiques et l'exploitation des RSN dans le projet de *Radar social*.

Automatisation du codage des données médiatiques

La prédiction automatisée de crises exploite des « données d'événement » (« *event data* ») constituées à partir de ce qui est et a été diffusé dans les médias transnationaux, régionaux et locaux (en presse écrite, radio et télévision). L'analyse des contenus médiatiques doit permettre de saisir ce qui se passe dans différents pays (les « événements »). L'objectif est d'identifier les mutations des rapports de forces nationaux et subnationaux, de prédire et d'anticiper des « instabilités ». Avec l'insertion du ICEWS dans le HSCB, l'automatisation du codage de « données d'événements » serait passée à une autre échelle (Schrodt, Van Bracke, 2013). Les bases de données porteraient sur un nombre de pays plus important et exploiteraient bien davantage d'articles ou de reportages audiovisuels (175 pays et 20 millions de nouvelles en 2012) que les versions du même type précédemment mises en œuvre. La vitesse du codage automatisé aurait également sensiblement augmenté, au point de produire des données d'événements « presque en temps réel » (*Ibid.*, p. 25). Cette accélération permettrait de réduire au minimum le délai entre, d'une part, la fabrique des « *event data* » sur l'évolution des rapports politiques dans le monde et, d'autre part, la manifestation *in situ* de ces rapports quotidiennement relayés par les médias.

Volume de données et vitesse de traitement sont mis en avant par les artisans des programmes à l'adresse des politiques (*Ibid.* ; O'Brien, 2010). En valorisant la portée et l'efficacité de ces technologies, il s'agit d'indiquer aux gouvernants américains qu'ils pourront optimiser leurs prises de décision. Dans la présentation de Philipp Schrodt et David Van Bracke (2013), un élément de ce travail de valorisation mérite une attention particulière. Les auteurs comparent à plusieurs reprises les performances de leurs systèmes aux capacités humaines, à leurs yeux sensiblement plus limitées. Cette comparaison peut être interprétée comme un travail de distinction élective dans le cadre d'une compétition intra-sectorielle. En effet, dans le secteur de la prédiction de crise, deux types de professionnels s'opposent au regard de leurs savoir-faire : le « *computational social scientist* » (le chercheur en sciences sociales computationnelles) et le « *political scientist* » (le politiste). Or, l'insistance avec laquelle les premiers soulignent les limites humaines des seconds (les politistes non-relégués par ordinateur) revient à disqualifier leurs compétences (estimées désormais obsolètes) et, ainsi, à délégitimer ces professionnels dans leurs prétentions à contribuer à la prise de décision politique.

Exploitation des RSN dans le projet de Radar social

Parmi les programmes développés dans le cadre du HSCB, le Radar social (« *Social radar* ») a été conçu comme un instrument de radiographie des mondes sociaux. Son système traite des données sur les émotions, les sentiments et les attitudes des populations civiles. L'ambition de ses concepteurs est de « détecter des signatures de comportements socioculturels » (HBSC, p. 47) et, par ce moyen, de prédire et d'anticiper l'émergence de conflits.

Ce « radar » exploite des gisements de *data* issues des médias (en utilisant les technologies de codage précédemment mentionnées) et des *data* extraites des réseaux sociaux numériques. Dans une présentation du projet de 2012, l'exploitation des RSN est davantage mise en exergue que dans la première présentation du projet datée de 2010 (Maybury, 2010). Cette attention est justifiée par la

généralisation de leurs usages et l'ampleur des mobilisations qui s'y manifestent, les auteurs faisant explicitement référence aux mouvements protestataires du « printemps arabe » de 2011 et 2012 (Costa, Boiney, 2012). Or, ces mouvements n'ont pas été anticipés par l'administration de Barack Obama, alors même qu'ils ont contribué à reconfigurer les *statu quo* géopolitiques en Afrique du nord et au Moyen-Orient. Aussi, dans le but de mieux détecter et prévoir de tels mouvements civils, des outils destinés à faire émerger les principaux sujets dont discutent les internautes ont été intégrés à la seconde version du Radar social (2012), principalement en associant l'analyse de conversation et l'analyse de sentiments grâce aux techniques d'« *opinion mining* » (Shellman, Covington, Zangrilli, 2014). Les logiciels du Radar visent en effet à cartographier des « constellations de sentiments-cibles » (Costa, Boiney, 2012, p. 7) et, par géolocalisation, à associer ces données à des pays ou des régions. Le système doit permettre d'« identifier des points de rupture qui signalent des changements majeurs de sentiments susceptibles d'avoir des effets sur le comportement des populations ou des gouvernements » (*Ibid.*, p. 2). La détection de ces ruptures, et la prédiction des changements politiques conséquents, permettent en retour de mettre en œuvre des tactiques de propagande, de cyber-déception et des opérations psychologiques adaptées à ces évolutions. Elles entrent à ce titre dans la prise de décision politique.

Le Radar social poursuit les mêmes objectifs que les machines prédictives mises en œuvre dans les années précédentes : prédire les conflits. Cependant, ce à partir de quoi la prédiction automatisée est réalisée a sensiblement évolué avec ce programme et, plus généralement, avec l'insertion du ICEWS dans le HSCB. Dans la lignée des programmes développés dans les années 1990, le projet *Senturion* (voir *supra*), par exemple, était essentiellement focalisé sur les décideurs politiques. Le Radar social a changé le point de focale. Le système porte sur les manifestations de soutien ou d'aversion des populations civiles vis-à-vis de ces décideurs. La recherche en modélisation computationnelle des comportements socioculturels organisée dans le cadre du HSCB est dévolue à ce déplacement d'objet. Les ingénieries algorithmiques sont développées pour automatiser l'indentification et la prédiction des liens entre gouvernés et gouvernants, entre suiveurs et *leaders* et, potentiellement, renforcer ou saper ce lien. Dans ce cas, les machines prédictives sont employées à la (re)configuration des environnements stratégiques civils dans lesquels la puissance américaine doit opérer.

Economie de la promesse (non tenue)

On se propose dans cette dernière partie de reconsidérer chaque dispositif sociotechnique précédemment présenté et de restituer leur développement dans l'« économie d'une promesse » (Joly, 2010). Il s'agira de montrer, dans un premier temps, en quoi la promesse d'une optimisation de la prise de décision politique grâce aux machines détectivo-prédictives n'est ni tenue, ni tenable. On se proposera, dans un second temps, d'interpréter le maintien de tels dispositifs malgré les déficiences de leur fonctionnement.

A suivre leurs artisans et leurs promoteurs, les technologies employées pour guider les drones vers leurs cibles permettraient de tuer exclusivement l'ennemi en le distinguant des civils innocents. Cependant, l'intensification du recours aux drones tueurs entre 2009 et 2013 n'a pas atteint ce but. Plusieurs études empiriques ont mis en avant la récurrence d'erreurs de ciblage causant la mort de civils sans lien avec les « terroristes » ou les « insurgés » (Aslam, 2014 ; Gil, 2014). Ces erreurs peuvent s'expliquer, d'abord, par la faiblesse des optiques utilisées sur les drones en altitude : les caméras embarquées ne permettent pas de distinguer distinctement les individus au sol. Elles peuvent s'expliquer, ensuite, par le fonctionnement de systèmes de détection qui associe un changement de comportement à une « déviance ». Afin de rendre évidents les travers de ce système, considérons l'exemple suivant : un homme en Irak n'est pas sympathisant de Daech mais son frère est combattant dans cette organisation. Ces deux frères ne se parlent plus depuis un certain temps. Leur mère décède. Pour organiser les funérailles ou pour se consoler mutuellement, le premier frère appelle régulièrement le second et le voit à plusieurs reprises. Ce changement le fera apparaître automatiquement dans le réseau « terroriste », comme un point nodal sur les graphes du contre-terrorisme. Dès lors, son exécution devient possible, voire nécessaire. Quoi qu'il en soit, les erreurs de ciblages et la mort de civils « innocents » créent de puissants ressentiments au sein des populations

autochtones. Des ressentiments qui alimentent les sentiments hostiles vis-à-vis des Etats-Unis et qui nourrissent des processus dits de « radicalisation » (*Ibid.*). La promesse de l'optimisation de la violence armée « légitime » n'a pas donc pas été tenue, pourtant le recours aux machines détectives dans la guerre à distance a été maintenu dans les années 2010.

Avec l'automatisation du codage des « données d'événement » développée dans les années 2000, une autre promesse a été faite : celle d'une transparence de l'état événementiel du monde qui permettrait d'objectiver avec précision les rapports de force nationaux et subnationaux. Le codage automatisé « presque en temps réel » devait permettre d'optimiser la prise de décision politique en la dotant d'une nouvelle ingénierie prédictive. Cependant, la mise en représentation de ce monde au moyen d'« *event data* » issues des médias pose un certain nombre de difficultés qui ne sont pas considérées par les ingénieurs du codage. Lorsque ces ingénieurs envisagent les limites de leur système, la raison invoquée est que « *les nouvelles ne sont [...] qu'une fraction étroite de tous les événements qui ont lieu quotidiennement* » (Schrodt, Van Bracke, 2013, p. 25). L'argument porte sur le fait que la couverture médiatique des évènements est partielle. La partialité de cette couverture dans la « fraction étroite » n'est pas quant à elle appréhendée. Pourtant, les travaux en sociologie du journalisme et en économie politique critique de l'information ont, de longue date, déconstruit l'idéal d'une représentation médiatique exacte et transparente du monde. Ces travaux ont montré notamment que le rôle des élites ou des grandes agences de presse internationales dans le fonctionnement des systèmes médiatiques a des effets conséquents quant à la *sélection* par les journalistes des événements et quant à leur *construction* (Stuart Hall et *al.*, 1978 ; Mattelart, 2016, 2017). Le projet d'identifier les rapports de forces nationaux et subnationaux à partir d'une analyse automatisée des médias se heurte à un autre obstacle. Dans de nombreux pays, les médias font l'objet d'un puissant système de contrôle. Or, ce contrôle permet de fabriquer des événements en conformité avec les lignes de la propagande nationale. Les tutelles sur les médias tunisiens sous le régime de Zine el-Abidine Ben Ali (1987-2011), par exemple, visaient (entre autre) à produire les représentations d'un pays stable, en mesure de rassurer les investisseurs internationaux. Aussi est-il bien hasardeux, au moyen du codage de nouvelles fabriquées dans ce type de dispositif de contrôle, de prédire quelque « instabilité » à venir.

En exploitant des données extraites des RSN, le Radar social est confronté à des difficultés similaires. Pour rappel, le Radar a été conçu pour prédire des révoltes et pour anticiper leurs conséquences géopolitiques. Afin de détecter en ligne des mouvements comme ceux du « printemps arabe », la version de 2012 du programme a intégré des outils d'analyse de conversation et d'« *opinion mining* ». Mais ces modifications ne changent rien quant à l'incapacité du système à remplir ses fonctions de prédiction. Le Radar social est inopérant parce que sa conception est emprunte d'un déterminisme technologique qui façonne une lecture erronée de l'histoire de ces mobilisations protestataires. En effet, si on prend pour exemple les révoltes en Tunisie et en Egypte en 2011, celles-ci ont commencé dans des régions intérieures déshéritées. Les réseaux sociaux numériques ont été l'une des arènes fragmentées où ces protestations se sont exprimées, mais les « révolutions » étaient déjà engagées. Il était donc impossible de prédire leur avènement à partir de ces réseaux. En somme, dans sa version de 2012, le Radar social a été façonné par la mythologie des « révolutions Facebook » ou « révolutions 2.0 », des révolutions qui n'ont pas eu lieu, et dont le principal travers consiste à surdéterminer le rôle qu'ont joué les réseaux sociaux dans la chute du régime de Zine el-Abidine Ben Ali ou d'Hosni Moubarak (Ferjani, Mattelart, 2011).

Les dispositifs de détection et de prédiction automatisées ne sont donc pas en mesure de tenir la promesse de leurs artisans-concepteurs, cependant chaque programme précité a été maintenu dans les années 2010. Plusieurs éléments peuvent être avancés pour interpréter la perpétuation de cette contradiction. Le travail promotionnel sur la performance des dispositifs, tout d'abord. Dans la littérature en Recherche et Développement appliquée aux usages militaires, des chiffres relatifs à la quantification de cette performance sont régulièrement avancés. La prédiction de crises des programmes développés dans le cadre de l'ICEWS, par exemple, est estimée fiable à 70% (O'Brien, 2010 ; Schrodt, Van Bracke, 2013). Cependant, ce chiffre n'est appuyé par aucune preuve et les conflits « prédits » ne sont pas mentionnés. La précision quantifiée des machines détective est aussi fréquemment mise en avant, mais elle n'est pas balancée par les chiffres concernant les victimes de leur imprécision. Quoi qu'il en soit, ces chiffres ne sont pas réellement en mesure de « duper » les

destinataires des services de détection et de prédiction. Aussi faut-il chercher le maintien de la contradiction à un autre niveau.

Derrière les promesses des ingénieurs se sont alignés les acteurs politiques et des capitaux importants. Cet alignement a eu un effet d'enrôlement sur d'autres secteurs dont on se bornera, ici, à indiquer quelques temps forts. Sous le mandat de Georges W. Bush, l'Etat fédéral a financé la recherche en *data mining*[6]. Au sein du complexe militaro-industriel —dans le département de la Défense et chez les prestataires privés— les filières ont été réorganisées pour exploiter des *data* « socioculturelles » à des fins sécuritaires (Gonzales, 2010, 2015). De ce point de vue, l'administration de Barack Obama n'a pas marqué de rupture avec celle de son prédécesseur. Au contraire. Sous sa présidence, les *data* et leur traitement pour « gouverner à travers le pouvoir civil [7]» ont été placés au centre des réformes stratégiques en matière de diplomatie et de développement. Le *21st Century Statecraft* en 2009 a officiellement acté la volonté de cette administration de mettre à niveau les outils de la politique étrangère américaine en tirant profit des « réseaux » et d'un « monde interconnecté »[8]. A ces fins, l'État fédéral a financé la recherche en sciences du comportement appliquée au traitement des *data*. Ainsi, les acteurs de différents secteurs (industriel, politique, militaire, académique) se sont solidarisés sur l'horizon d'attente des promesses des dispositifs « *Big Data* et algorithme ». Dès lors, le maintien de programmes qui ne peuvent fonctionner à la hauteur de l'exactitude escomptée n'est plus si contradictoire qu'il y paraissait au premier abord. L'horizon d'attente n'appelle pas l'abandon de ces programmes mais, au contraire, leur perfectionnement. Les failles de la technologie sont appelées à être compensées par de nouvelles innovations tant que celles-ci concourent à la réalisation de la promesse. La prédiction de crise illustre cette logique : alors qu'elle est dite fiable à 70% (sans preuve tangible), l'investissement est pourtant porté sur la correction des 30% restants.

Conclusion

On a œuvré dans cet article à analyser l'intégration des dispositifs « *Big Data* et algorithmes » dans le secteur militaire au tournant de la guerre « centrée sur la culture ». Il ressort de cette investigation que les promesses de la détection et de la prédiction automatisées, au service de l'optimisation de la prise de décision politique, ont gouverné cette intégration. Nous avons souligné l'inefficacité des programmes mis en œuvre au regard des finalités que leur assignent leurs artisans-concepteurs. Afin de rendre compte de cette contradiction apparente, ont été esquissés les effets d'enrôlement entre acteurs de différents secteurs dans une « économie de la promesse » agencée au plus haut niveau de l'Etat fédéral. En guise de conclusion, on souhaiterait souligner que l'automatisation de la détection et de la prédiction *doivent* fonctionner avec des marges d'erreur. Ce déplacement de perspective implique de rappeler que l'incertitude est une condition du pouvoir de l'agir politique. En effet, si un homme politique n'a plus à interpréter, s'il doit exécuter ce que prédisent les machines, alors sa marge d'action est réduite à sa partie congrue. Plus les prédictions sont exactes, moins sa marge de manœuvre est importante. Autrement dit, la faillite des machines détectivo-prédictives est une condition de survie des professionnels de la politique. Raison pour laquelle, probablement, les politiques font preuve de tant de complaisance vis-à-vis des approximations des oracles mécanisés.

. . . . . .

[6] Cf. United States Genaral Accounting Office, « Data mining. Federal efforts cover a wide range of uses », May 2004.

[7] Cf. U.S. Departement of state, « Leading through civilian power : 2010 quadrennial diplomacy and developpement review », [en ligne], consulté le 19 février 2017, https://www.state.gov/documents/organization/153108.pdf

[8] Cf. U.S. Departement of state, « 21st Century statecraft », [en ligne], consulté le 26 mars 2018, https://2009-2017.state.gov/statecraft/overview/index.htm

Références bibliographiques

Abdollahian, Mark ; Baranick, Michael ; Efird, Brian et Kugler, Jacek (2006), *Senturion: a predictive political simulation model*, Center for technology and national security policy national defense university, [en ligne], Consulté 19 septembre 2017, http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA454175&Location=U2&doc GetTRDoc.pdf

Andriole, Stephen. J. et Young, Robert A. (1977), « Toward the development of an Integrated Crisis Warning System », *International Studies Quarterly*, n°21, p. 107-150.

Aslam, Walid (2014), « Terrorist relocation and the societal consequences of US drone strikes in Pakistan », *The remote controle digest*, [en ligne], Consulté le 20 mars 2017, http://www.oxfordresearchgroup.org.uk/sites/default/files/Remote%20Control%20Digest.pdf

Chamayou, Grégoire (2016), *Théorie du drone*, Paris : La fabrique.

Colonomos, Ariel (2014), *La politique des oracles. Raconter le futur aujourd'hui*, Paris : Albin Michel (collection « Bibliothèque des Idées »).

Costa, Barry et Boiney, John (2012), *Social radar*, Mitre Corporation, [en ligne], consultée le 26 juin 2017, https://www.mitre.org/sites/default/files/pdf/12_0581.pdf

Ferjani, Riadh, Mattelart, Tristan (2011), « Monde arabe : les révolutions 2.0 n'ont pas eu lieu », *Médias*, n° 30, p. 67-94.

Forget, François (2001), « Réseaux d'informations et mutations stratégiques », *Panoramiques*, p. 65-78.

Gill, Paul (2014), « The impact of drone attacks on terrorism : the case of Pakistan », *The remote controle digest*, [en ligne], Consulté le 20 mars 2017, http://www.oxfordresearchgroup.org.uk/sites/default/files/Remote%20Control%20Digest.pdf

Gonzales, Roberto J. (2010), *Militarizing Culture*, Walnut creek : Left coast press.

Gonzales, Roberto J. (2015), « Seeing into hearts and minds, Part 1 », *Anthropology today*, n° 31, p. 8-18.

Hall, Stuart ; Critcher, Chas ; Jefferson, Tony, Clarke, Jhon et Roberts, Brian (19878), *Policing the crisis : Mugging, the State, and Law and Order*, Londres : Palgram Macmillan.

Hopkins, Benjamin D. (2016), « The longue durée of Human Terrain : politics, cultural knowledge and the technical fix », *Anthropology today*, n°32, p. 8-12.

Joly, Pierre-Benoit (2010), « On the economics of technoscientific promises » (p. 203-222), in Akrich, Madelein ; Barthe Yannick ; Muniesa, Fabian ; Mustar, Philippe (dir.), *Débordements : mélanges offerts à Michel Callon*, Paris : Presse des Mines.

Kipp, Jacob ; Grau, Lester ; Prinslow, Karl et Smith, Don (2006), « The Human Terrain System : a CORDS for the 21st Century », *Military Review*, n° 86, p. 8-15.

Mattelart, Armand ; Vitalis, André (2014), *Le profilage des populations, du livret ouvrier au cybercontrôle*, Paris : La découverte.

Mattelart Tristan (2016), "Déconstruire l'argument de la diversité de l'information à l'heure du numérique : le cas des nouvelles internationales", *Les Enjeux de l'Information et de la Communication*, n°17/2, 2016, p.113 à126, consulté le lundi 5 novembre 2018, [en ligne] URL : https://lesenjeux.univ-grenoble-alpes.fr/2016-dossier/07-Mattelart/

Mattelart, Tristan (2017), « Les enjeux de la circulation transnationale de l'information : des agences de presse aux plateformes du web », in Koch, Olivier et Mattelart, Tristan (dir.), Géopolitique des télévisions transnationales d'information, Paris : Mare & Martin.

Maybury, Mark (2010), « Social radar for smart power » (p. 26-36), in Schmorrow, Dylan et Nicholson, Denise (éd.), *Cross-cultural decision making*, New York : Taylor & Francis group.

O'Brien, Sean (2002), « Anticipating the good, the bad, and the ugly : an early warning approach to conflict and instability analysis », *The journal of conflict resolution*, vol. 46, n° 6, p. 791-811.

O'Brien, Sean (2010), « Crisis early warning and decision support : contemporary approaches and thoughts on future research », *International studies review*, vol. 12, n°1, p. 87-104.

Pucheu, David (2017), « Social Big Data. Le phantasme d'une nouvelle physique sociale », *Etudes digitales*, vol. 2, n°2, p. 89-106.

Schrodt, Philipp A. ; Van Bracke, David (2013), « Automated coding political event data » (p. 23-49), in Subralmanian, V.S. (éd.), *Handbook of computational approaches to counterterrorism*, New York : Springer science.

Shellman, Steve ; Covington, Michael et Zangrilli, Marcia (2014), « Sentiment & discourse analysis : theory, extraction, and application », *Socio-Cultural Analysis with the Reconnaissance, Surveillance, and Intelligence Paradigm*, [en ligne], Consulté le 13 février 2018, http://nsiteam.com/social/wp-content/uploads/2016/01/Socio-Cultural-Analysis-with-the-Reconnaisance-Surveillance-and-Intelligence-Paradigm.pdf

Singer, Peter (2009), *Wired for war. The robotics revolution and conflict in the 21st Century*, New York : Penguin books.

Le journalisme saisi par les *Big Data* ? Résistances épistémologiques, ruptures économiques et adaptations professionnelles

Are Big Data invading Journalism? Epistemological Resistances, economic Breakings and professional adaptations

¿Está tocado por el Big Data el periodismo? Resistencias epistemológicas, rupturas económicas y adaptaciones profesionales

Article inédit, mis en ligne le 15 novembre 2018.

Alexandre JOUX

Alexandre Joux est maître de conférences en Sciences de l'information et de la communication, chercheur à l'IMSIC EA 7492 et directeur de l'École de journalisme et de communication d'Aix-Marseille (EJCAM). Ses recherches portent sur l'économie de l'information et des industries culturelles, en particulier dans les contextes de convergence et d'internationalisation, et sur les évolutions du journalisme dans son rapport aux environnements médiatiques et numériques.

alexandre.joux@univ-amu.fr

Marc BASSONI

Marc Bassoni est maître de conférences (HDR) en Sciences de l'information et de la communication, chercheur à l'IMSIC EA 7492 et directeur des études à l'École de journalisme et de communication d'Aix-Marseille (EJCAM). Ses recherches portent sur l'économie des médias et l'évolution des pratiques éditoriales et journalistiques à l'aune de la révolution numérique.

marc.bassoni@univ-amu.fr

Résumé

La déferlante des données massives (*big data*) interroge le journalisme contemporain, ses pratiques, mais aussi les modèles économiques des médias et l'organisation des rédactions. Les enjeux sont protéiformes. Nous mettons tout d'abord en évidence la résistance épistémologique du métier face aux promesses d'un journalisme intégralement robotisé. En effet, le data journalisme se réfère toujours à un terrain identifié auquel les données doivent renvoyer à des fins de validation. Nous montrons ensuite que l'usage des données massives permet aux médias non seulement de personnaliser plus finement l'offre d'information, mais également de favoriser un processus d'innovation qui tend à hybrider de plus en plus les savoir-faire journalistiques avec des compétences techniques extérieures au métier. A défaut de disparaître, le métier mute en se recentrant sur certains de ses fondamentaux.

Mots clés

Journalisme automatisé, data journalisme, *big data*, épistémologie, marchandisation de l'information.

Abstract

Faced with the deluge of massive amounts of data (*big data*) today's journalism wonders about its practices, and also about the economic models of the media and the organization of editorial teams. The issues are multifaceted. At first we emphasize the epistemological reluctance of the profession to the promises of totally automated journalism. Actually data journalism is always linked with an identified ground to which data must refer in order to validate it. Then we show that the massive use of data enables the media not only to personalize their offer of information more precisely, but also to promote an innovation process which tends to combine the know-how of journalism with technical skills unrelated to the journalism professions. Even if the profession does not disappear, it undergoes an evolution by refocusing on some of its fundamental principles.

Key words

Automated journalism, data journalism, big data, epistemology, information commodification.

Resumen

La invasión de los largos datos (*big data*) no sólo hace reflexionar el periodismo contemporáneo, sobre sus prácticas, sino también sobre los modelos económicos de los medias y la organización de las redacciones. En primer lugar ponemos de relieve la resistencia epistemológica de la profesión frente a las promesas de un periodismo totalmente robotizado. En efecto, siempre se refiere el periodismo de datos a un terreno identificado al que los datos deben remitir como forma de validación. Demostramos luego que el uso de los datos masivos no sólo permite a los medios de comunicación personalizar con un análisis más fino la oferta de información, sino también fomentar el proceso de innovación que tiene por objeto hibridar cada vez más los conocimientos periodísticos con conocimientos especializados externos a la profesión. Si no desaparece, muta la profesión centrándose en algunos de sus fundamentos.

Palabras claves

Periodismo automatizado, data periodismo, largos datos, epistemología, mercantilización de la información.

Introduction

Les médias d'information sont parmi les premiers concernés par la déferlante nouvelle de données. Celle-ci interroge le journalisme professionnel, ses pratiques, mais aussi les modèles économiques des médias et l'organisation des rédactions. Il n'y a alors qu'un pas pour faire des *big data* le levier d'une transformation en profondeur du journalisme.

En l'absence de définition partagée de la notion, nous retiendrons ici par *big data* l'ensemble des données qui relèvent de la règle des trois V souvent invoquée (le volume, la variété, la vélocité dans la récolte), avec comme discours d'accompagnement l'idée que les *big data* autorisent une forme d'intelligence nouvelle ayant l'aura de la vérité et de l'objectivité (Boyd, Crawford, 2012). Ce discours atteste de la dimension prescriptrice des *big data*, Dominique Cardon proposant à cet égard d'ajouter un quatrième V à la règle de trois, à savoir les valeurs véhiculées par les données (Cardon, 2015). La

confrontation des *big data* à d'autres méthodes de saisie du réel par les données est de ce point de vue révélatrice des enjeux et des limites de la notion, ce dont atteste le rapprochement trompeur, mais souvent opéré, entre data journalisme et *big data*.

Bien loin d'être une adaptation du journalisme aux *big data*, le data journalisme rappelle que le journalisme se réfère toujours à un terrain identifié auquel la donnée doit renvoyer *in fine* afin de pouvoir être vérifiée, trahissant de ce point de vue la résistance épistémologique du journalisme face aux promesses des *big data*. Cette résistance s'accompagne toutefois d'une adaptation des rédactions. Les algorithmes associés aux masses de données peuvent être utilisés autrement, parce qu'ils produisent des signaux faibles permettant d'explorer des sujets que la rédaction aurait pu manquer, ou encore parce qu'ils permettent d'automatiser le traitement de certaines informations, libérant ainsi les journalistes de tâches à faible valeur ajoutée.

Par ailleurs, les masses de données sont de plus en plus exploitées par certaines rédactions qui les utilisent comme des leviers puissants de marketing, rendant possible une personnalisation de la publicité, voire une personnalisation de l'offre d'information. Cette dernière possibilité vient transgresser les frontières professionnelles du journalisme qui séparent la rédaction des fonctions commerciales des journaux. La personnalisation de l'offre d'information peut ainsi conduire à un journalisme tiré par la demande qui vient mettre à mal le rôle d'intermédiation des médias d'information, tout en redéfinissant leurs arbitrages éditoriaux. Enfin, l'invocation des *big data* au sein des rédactions témoigne d'un impératif d'innovation qui fait émerger des pratiques aux frontières du journalisme professionnel. Si ces frontières mouvantes ont toujours été l'enjeu d'une adaptation négociée à un nouvel environnement économique, technique et social (Ruellan, 1992), elles sont aussi, pour le journalisme, un moyen de gérer le renouvellement de ses profils en accueillant des compétences nouvelles et, pour les firmes médiatiques, un moyen de repenser leur organisation.

Nous nous proposons ici de mettre en perspective ces enjeux dans une démarche d'exploration scientifique ancrée dans les SIC, en mobilisant une approche transdisciplinaire du journalisme, sociologique et économique. En termes méthodologiques, nous nous appuyons sur une revue de la littérature scientifique, avec une cartographie des premières recherches initiées sur le rapport du journalisme aux *big data*, principalement dans un contexte anglo-saxon (Lewis, Westlund, 2015). Des témoignages d'acteurs sont également mobilisés (entretien avec Nicolas Kayser-Bril réalisé en janvier 2017 ; atelier avec David Dieudonné, Directeur du *Google News Lab France*, en mars 2017).

Résistances épistémologiques à l'automatisation du journalisme

S'interroger sur le rôle des *big data* et des algorithmes dans le champ journalistique conduit à s'intéresser au journalisme automatisé (ou « robot-journalisme »), à savoir la production d'articles par des logiciels à partir du traitement algorithmique de grandes masses de données. Le journalisme robotisé semble en effet s'imposer comme un modèle d'« objectivité » poussant à son paroxysme l'exigence de neutralité du journaliste dans l'observation du réel grâce à la suppression de l'intervention humaine. Souvent donné en exemple, le *Los Angeles Times* publie ainsi depuis 2014 des articles sur l'activité sismique en Californie grâce au logiciel-robot *Quakebot* qui exploite les données de l'*U.S. Geological Survey*. Si la perspective de rédactions sans journalistes inquiète, ce journalisme automatisé s'accompagne aussi de promesses. Parmi ces promesses, on notera la possibilité de multiplier les articles pour répondre à des demandes de niche et pour traiter de sujets que les rédactions n'avaient pas les moyens de couvrir jusqu'alors (voir Carlson, 2015). Cette prolixité des robots-journalistes doit dès lors libérer les journalistes des tâches répétitives pour se consacrer à des enquêtes ou des projets qu'une intelligence artificielle ne pourra pas assumer.

Ces promesses font toutefois l'objet de résistances de la part des journalistes que le chercheur américain Matt Carlson a recensées dans les débats qui ont accompagné le lancement de *Narrative Science* aux Etats-Unis, une entreprise qui propose du journalisme automatisé mais qui ne se revendique pas comme une entreprise de journalisme (Carlson, 2015). Parmi ces résistances, la première d'entre elles consiste à questionner le renversement des rapports entre l'offre et la

demande dans la production de l'information. Le journalisme se définirait d'abord par l'offre, par les choix de couverture qu'une rédaction opère, c'est-à-dire la sélection de l'information qui mérite d'être portée à la connaissance du public, ce que les anglo-saxons qualifient de « *newsworthiness* ». Cette sélection de l'information est au cœur de l'autorité journalistique qui assume ici son rôle de « *gate keeping* ». *A contrario*, les tenants des *big data* vont dénoncer les insuffisances de l'intermédiation journalistique, avec ses choix humains, auxquels les algorithmes opposent une connaissance plus fine des besoins des utilisateurs, la « pertinence » ou « *relevance* » (Gillespie, 2014) de leurs réponses se substituant ici aux logiques éditoriales. De ce point de vue, les *big data* constitueraient l'aboutissement d'un processus qui, depuis les années 1990, donne à la demande une plus grande importance dans les choix éditoriaux. Mais elles en transfigurent la signification : après l'impératif publicitaire et ses *metrics* (Ouakrat, 2012), la demande s'imposerait au nom de la *relevance* des informations sélectionnées, c'est-à-dire au nom d'une revendication épistémologique qui ferait des algorithmes le vecteur d'une nouvelle science des publics et de leurs attentes.

Les divergences portent également sur le statut des faits. Ces derniers, parce qu'ils s'inscrivent dans une réalité complexe, doivent de plus en plus être contextualisés et interprétés, ce qui interdit de produire un compte-rendu épuré du réel, faisant de l'objectivité une norme à vocation d'abord régulatrice (Schudson, 2001). A cette interprétation journalistique des faits, suspectée de laisser place au risque de biais idéologique, le journalisme automatisé de *Narrative Science* va opposer la possibilité de neutraliser toute forme de subjectivité dans le traitement de l'information grâce à l'application de processus normés. A l'évidence, des tensions épistémologiques se manifestent entre, d'une part, l'expertise journalistique et son rapport négocié au réel et, d'autre part, la prétention des algorithmes à faire émerger de manière normée des faits à partir du traitement de gigantesques bases de données.

Afin d'explorer ces tensions, nous nous proposons d'interroger ici le rapport du data journalisme aux données parce que cette nouvelle manière de faire du journalisme, qui émerge à la fin des années 2000, revendique justement la possibilité, à partir des données, de produire des contenus qui s'inscrivent au cœur du champ journalistique (Croissant, Touboul, 2013), sans lui être extérieur comme l'est le journalisme robotisé de *Narrative Science*. La démarche initiée par le data journalisme, ou journalisme de données, peut alors se comprendre comme une manière d'apprivoiser la profusion nouvelle de données.

Le data journalisme accorde une importance majeure aux données et à leur exploitation, prenant ainsi ses distances avec le « terrain », au moins en première intention. Il a témoigné, aux Etats-Unis, d'une évolution du journalisme qui, après avoir favorisé l'observation sur le terrain et les interviews dès le 19ème siècle, sur le modèle des sciences sociales, revient aux documents, donc aux données et à leurs archives, pour explorer des tendances sur le temps long (Anderson, 2015). Ce retour aux documents ne signifie pas pour autant une prise de distance à l'égard des canons scientifiques des sciences sociales, qui restent fortement mobilisés dans les discours des journalistes quand ces derniers cherchent à légitimer leur prétention à l'objectivité. Ainsi, l'application, dans les années 1970, des méthodes des sciences sociales au « journalisme de précision » (Meyer, 1973) imposera dans le champ journalistique le recours à l'informatique et aux statistiques comme moyen d'établir des faits. L'importance de l'informatique dans le journalisme se retrouve aujourd'hui dans le journalisme robotisé, dans le journalisme computationnel, lequel donne la priorité aux algorithmes sur les savoir-faire journalistiques, ou encore dans le data journalisme (Coddington, 2015). Ce dernier est à part ; il revendique en effet une dimension hybride ; il associe aux journalistes-rédacteurs des compétences venues du *hacking* (Dagiral, Parasie, 2011) et fait également appel aux destinataires de l'information (public) qui pourront ainsi contribuer à la qualification des données, le cas d'école restant en la matière le dispositif du *Guardian* pour les notes de frais des députés britanniques (Daniel, Flew, 2010). A cet égard, le statut de la donnée, ainsi que la manière dont celle-ci a été qualifiée, sont essentiels dans le data journalisme.

D'abord présentées comme des marqueurs de vérité parce qu'elles échapperaient à toute interprétation subjective, les données vont être ensuite repensées par le data journalisme pour mieux prendre en compte les conditions de leur établissement, leur structure, et donc aussi les possibilités de leur exploitation. Le data journalisme va alors s'engager dans une démarche constructiviste qui ne

prétend pas saisir le réel à l'état pur à travers la donnée pure (*raw data*). L'un des actes fondateurs du data journalisme est la déclaration d'Adrian Holovaty qui, dès 2006, exigeait une forme nouvelle de journalisme : « *Par exemple, supposons qu'un journal ait écrit un article sur un incendie local. [...] ce que je veux vraiment pouvoir faire, c'est explorer les faits bruts de cette histoire un par un, avec des couches d'attribution et une infrastructure permettant de comparer les détails de l'incendie avec ceux d'incendies précédents : date, heure, lieu, victimes, numéro de la caserne de pompiers, distance de la caserne, nom et nombre d'années d'expérience de chaque pompier présent sur les lieux, temps mis par les pompiers pour arriver sur place, et les incendies ultérieurs, le cas échant.* » (Kayser-Bril et *alii*, 2013, p. 24). Or Nicolas Kayser-Bril, qui est à l'origine de l'édition française du *Guide du data journalisme* d'où la citation a été tirée, indique que cette vision d'Adrian Holovaty est pertinente uniquement parce qu'il s'agit dans son exemple d'une information hyper-locale. Elle ne nécessite pas de contextualisation particulière. Pour Nicolas Kayser Bril, la donnée seule n'est pas signifiante si elle est mise à disposition du public sans l'histoire qui lui donne sens.

Par ailleurs, la structure des bases de données n'est pas neutre. Là encore, les projets d'Adrian Holovaty témoignent des limites du premier data journalisme. Celui-ci a accordé une place très importante aux jeux de données pour les porter en l'état à la connaissance du public, des dispositifs visuels permettant ensuite d'opérer des tris dans ces données pour en faire émerger des informations signifiantes. Ainsi, Adrian Holovaty a-t-il lancé dès 2005 le site *chicagocrime.org* qui superpose sur une carte *Google Maps* les données en *open data* récupérées auprès du *Chicago Police Department*. Le site offre une cartographie interactive des crimes commis à Chicago. Ce site a révélé leur fréquence sur des secteurs urbains auparavant peu traités par la presse. Le data journalisme permet alors de corriger « l'inégale distribution des sources sur le territoire urbain » (Parasie, Dagiral, 2013, p. 56) tout en libérant en partie les journalistes de leur dépendance aux sources institutionnelles.

Il y a ici une croyance dans la force de la donnée brute qu'une enquête du *Los Angeles Times* va révéler comme trop naïve. En effet, le quotidien de la Cité des Anges s'interroge sur la pertinence des données mises à disposition par *everyblock.com*, un autre site d'Adrian Holovaty qui a étendu à l'ensemble du territoire américain le modèle imaginé avec *chicagocrime.org*. Sur *everyblock.com*, le *City Hall* de Los Angeles apparaît comme l'endroit le plus criminogène de la ville, alors que les journalistes ne ressentent pas cette réalité sur ce terrain bien connu d'eux (puisqu'il s'agit de l'endroit où ils ont leurs bureaux !). Après enquête (Welsh, Smith, 2009), ils découvriront ainsi que le *Los Angeles Police Department* (*LAPD*), dont *everyblock.com* exploite les données, assigne par défaut le code postal du *City Hall* à tous les crimes dont la géolocalisation est problématique, ce qui provoque ce biais majeur sur *everyblock.com* et ce qui confirme la nécessité pour le data journaliste, après une phase d'exploitation des données, de retourner sur le terrain pour tester la pertinence des informations communiquées.

Cette approche s'oppose à la vision de Chris Anderson pour qui les *big data* changent en profondeur notre rapport à la vérité. Elles provoqueraient ainsi dans la science la même révolution copernicienne que celle opérée par *Google* dans la saisie statistique de la pertinence d'une page web, laquelle repose sur l'affectation de valeurs aux liens hypertextes (Cardon, 2013). Les *big data* et leur traitement mathématique rendraient caduques les anciennes méthodes scientifiques avec en ligne de mire la fin du dispositif empirique de vérification. D'où la sentence lapidaire de Chris Anderson : « *patabytes allow us to say : correlation is enough* » (Anderson, 2008). Comme dans la recherche en ligne, le critère d'efficacité du « *good enough* » l'emporte sur le test de réalité. La corrélation ne vaut pas preuve, elle « suffit ». Dès lors qu'elles échappent aux capacités de traitement humain et tendent vers l'exhaustivité, les masses de données rendraient obsolète le recours préalable au tri et au codage des informations au sein de bases structurées pour être interrogées. Or, justement, l'exemple des limites d'*everyblock.com* sur le positionnement interactif des crimes à Los Angeles montre la nécessité de maîtriser les modalités qui président à l'élaboration par le *LAPD* de la base de données des crimes à Los Angeles, sauf à énoncer des vérités qui n'en sont pas ! De ce point de vue, le data journalisme, parce qu'il maintient l'exigence de vérification, parce qu'il cherche donc à interpréter les données, et parce qu'il ne se résout pas au « *good enough* », relève d'une démarche épistémologique qui n'est pas celle des promoteurs des *big data*, au moins dans leurs promesses les plus radicales. Comme l'indique Dominique Cardon, les *big data* posent des problèmes opérationnels, plus que le « *good enough* » n'en résout, à savoir le problème des « corrélations sans causes » (Cardon, 2015, p.

21) et celui de la faiblesse du ratio « signal/bruit ». Le premier problème est un grand classique de l'analyse de données. Il ouvre la voie à une exubérance de micro-théories contingentes qui alimentent l'illusion de la fin des sciences de l'homme et de la société. Le second problème tient lui au fait que les données massives ainsi traitées offrent en fait une faible probabilité de présenter un vrai intérêt (signal) pour l'utilisateur.

Pour autant, il serait exagéré de considérer que les corrélations sans causes produites par des algorithmes opérant sur de grandes masses de données sont sans intérêt pour le data journalisme. Certes, le data journalisme favorise, comme le proclame Nicolas Kayser-Brill, les « bases de données structurées », celles qui sont questionnables parce qu'elles ont fait l'objet d'une construction maîtrisée par le journaliste, ou parce qu'elles sont suffisamment transparentes pour que le journaliste puisse exploiter les normes qui ont présidé à leur établissement. Il n'en demeure pas moins qu'il y a une utilité des données massives, celles qui excèdent par leur volume toute capacité d'appréhension et de tri, celles aussi pour lesquelles la transparence du tri algorithmique est inexistante. C'est ce que rappelle Eric Scherer quand il s'interroge sur l'utilisation des *big data* par les rédactions : « *Le tri algorithmique – qui gère les grands nombres et donc explique des choses qu'on ne peut pas voir – peut l'aider [le journaliste] à détecter des signaux faibles, à pointer vers des sujets où les rédactions restent frileuses* » (Scherer, 2015, p. 13).

Des contenus tirés « par la demande » qui émanent d'organisations médiatiques en voie de reconfiguration

L'inscription du journalisme au cœur d'un écosystème informationnel dominé par les « *data* » ne provoque pas seulement un débat sur le rapport entre données et vérité journalistique. Les *big data* s'invitent également dans les logiques marketing des entreprises d'information. Leur premier effet structurel est d'accélérer l'assujettissement de la production éditoriale à la perception des attentes du public. Le phénomène est déjà ancien ; le poids croissant des régies publicitaires avait, dès la fin des années 1980, fait basculer le journalisme traditionnel, bâti jadis selon une logique de l'offre, dans une logique de la demande. Ce basculement est désormais accéléré et amplifié. La demande, extrêmement labile, est plus prégnante que jamais. De nouvelles pratiques journalistiques attestent ce changement de rythme, en même temps qu'elles mettent en relief ses enjeux éditoriaux et économiques.

Au début des années 1990, John Mc Manus prophétisait l'assomption du journalisme tiré par la demande, ou « *Market*-Driven *Journalism* » (Mc Manus, 1994). Force est aujourd'hui de constater que la pratique répandue du « *Google* [ou *Facebook*]-journalisme » donne un écho saisissant à cette anticipation. Certains médias – *BuzzFeed*, par exemple, ou bien le site pour adolescents *Melty* – procèdent à une veille permanente *via Google Trends* et les réseaux sociaux (*hashtags* les plus discutés) pour déterminer, en temps réel, les questions qui font le « *buzz* » et préoccupent leurs destinataires. La plateforme *CrowdTangle*, rachetée en 2016 par *Facebook*, favorise par exemple le suivi analytique des contenus diffusés sur ces réseaux. Des algorithmes prédictifs peuvent même renforcer ce travail de veille en soulignant les sujets et les angles de couverture qui pourraient devenir viraux et assurer ainsi l'audience la plus large possible. En fonction de tels signaux perçus, ces médias agencent et reconfigurent au fil de l'eau leur offre d'information.

Autres pratiques journalistiques notables, celle de « l'information-service » et celle du journalisme dit « de solutions », ou journalisme « d'impact » (Gyldensted, 2015). Pratique ancienne, constitutive même de l'information de proximité, « l'information-service » est aujourd'hui totalement renouvelée à l'aune de la massification des usages du numérique. En tant qu'information utile pour le quotidien des destinataires, issue d'une source identifiée (extérieure au média diffuseur) et transmise au public sans retouche éditoriale de la part des journalistes, « l'information-service » fait désormais l'objet d'interactions avec les médias, interactions qui favorisent l'enrichissement des bases de données commerciales de ces derniers. Ces interactions permettent d'affiner la connaissance des « destinataires-clients » (*via* leurs déclarations, le recueil de leurs intentions, l'historique de leurs comportements tracés). Elles permettent *in fine* de qualifier ces bases de données en vue d'une

exploitation commerciale par d'autres services des médias concernés. Cette transversalité entre l'information et la communication commerciale est d'autant plus efficace qu'elle repose, comme en matière de « *Google*-journalisme », sur une appréhension fine des attentes et des intentions du public. C'est un enjeu du même ordre qui est associé au journalisme dit « de solutions ». Ce type de pratique, qui connait un fort engouement, favorise l'implication des destinataires de l'information au cœur même de la production journalistique. Ces destinataires sont non seulement des citoyens auxquels on vient proposer des « solutions » pour leur permettre d'affronter, à leur niveau, les grands défis de l'époque, mais également des pourvoyeurs de *data* monétisables au bénéfice d'activités diverses rassemblées sous une même marque-média (ou « marque-ombrelle »). Le journalisme dit « de solutions » renforce donc, au sein des médias, le tropisme en faveur du « marketing des traces » (Kessous, 2011) et de l'approfondissement de la relation de personnalisation autour de l'information produite.

L'accélération de ce basculement d'une logique éditoriale de l'offre vers une logique centrée désormais sur la demande a non seulement des effets sur les contenus produits, mais aussi sur l'économie des médias visés. Concernant la nature de l'information produite, l'effet collatéral est bien sûr un effet d'auto-renforcement. Une demande explicite suscite une offre qui, *a posteriori*, légitime et étaye cette demande au risque d'ailleurs d'éclipser d'autres sujets ou d'autres thématiques. La généralisation de telles pratiques « pro-demande » peut donc catalyser un appauvrissement qualitatif de cette partie de l'écosystème informationnel qui ne dépend que des recettes publicitaires. Cette dépendance exclusive oblige en effet les offreurs d'information à courtiser le même segment (large) du grand public et à proposer les mêmes sujets plébiscités, fédérateurs et non-clivants. On retrouve ici un résultat cohérent avec l'ancienne « loi de Steiner » (Bassoni, Joux, 2014, p. 86), loi selon laquelle la pluralité des offreurs ne rime pas forcément avec la diversité des contenus proposés au public.

Concernant l'économie des médias visés, ce basculement en faveur de la demande du public a des effets nombreux et structurants. Il catalyse en fait un reformatage en profondeur du modèle d'affaires de ces médias. Le premier effet à signaler a trait à la gestion de la manne publicitaire. Si les algorithmes permettent d'affiner la connaissance de l'audience, de ses caractéristiques et de ses attentes, ils renforcent par la même occasion la rentabilité liée à la monétisation de cette audience auprès des annonceurs. On appelle précisément publicité « programmatique » cette communication adaptée au public-cible et adossée à un traitement de données massives. Depuis 2016, celle-ci représente déjà plus de la moitié des investissements publicitaires en ligne en Europe (IAB, 2017). Les médias, légitimement soucieux de leur pérennité et de leur autonomie financière, doivent donc, plus que jamais, prendre en compte cette dimension économique cardinale de leur environnement. Cet enjeu éclaire d'ailleurs les ressorts de la bataille qui se joue actuellement entre les médias traditionnels et les GAFA au sujet du partage du « gâteau » publicitaire. On le sait, ces médias se plaignent des modalités d'un partage qu'ils jugent déséquilibré (Guibert et *alii*, 2016, p. 162). Mais bien au-delà des revenus à rétrocéder, le sujet le plus sensible concerne la maîtrise pleine et entière des données-clients. Avec son format *Instant Articles, Facebook* – nonobstant ses dénégations – n'a pas tout à fait renoncé à exploiter pour son propre compte les données relatives aux médias partenaires de la plateforme. Les projets d'alliances dans le domaine de la publicité numérique qui fleurissent actuellement entre médias traditionnels (en France, *Gravity* et *Skyline*) doivent être appréciés à l'aune de ce bras de fer ; ils marquent, de la part des médias, une vraie volonté de s'autonomiser vis-à-vis des poids lourds de l'économie numérique.

Un second effet économique est à signaler ; il concerne la tendance à personnaliser de plus en plus finement l'offre d'information et à favoriser une forte discrimination tarifaire. En tenant compte des centres d'intérêt de tout un chacun, de ses échanges personnels et des interactions qu'il alimente *via* les réseaux sociaux, les algorithmes peuvent éditorialiser/personnaliser à l'infini l'information qui lui est dédiée. Comme l'a prétendu Mark Zuckerberg, le 6 novembre 2014, le but est « *de bâtir le parfait journal personnalisé de tout le monde* », c'est-à-dire un quotidien sur mesure dédié à chaque internaute. L'enjeu commercial est double ici : il s'agit tout à la fois de fidéliser celui-ci et de l'enfermer au sein d'un écosystème cohérent et de l'exposer à une publicité ciblée toujours plus efficace. En individualisant ainsi à l'extrême la relation-client, la personnalisation des contenus autorise une discrimination tarifaire poussée. Comme l'indique l'économiste Jean-Pascal Gayant, « ce

que nous connaissons déjà pour le transport aérien ou ferroviaire a vocation à se développer pour les moindres objets et services » (Gayant, 2015), y compris les services informationnels. Cette possibilité, si elle est exploitée, ouvre donc la voie à une grande pluralité de modèles d'affaires au sein de l'univers médiatique. Subséquemment, les journalistes et les rédactions devront faire montre d'une grande adaptabilité aux contextes divers au sein desquels ils seront amenés à travailler. Cette adaptabilité va bien sûr de pair avec la reconfiguration des organisations médiatiques.

En matière de reconfiguration organisationnelle, deux niveaux différents doivent être distingués. Le premier concerne la rédaction *stricto sensu* ; le second a trait à la firme médiatique *lato sensu.* Au niveau de chaque rédaction, l'adaptabilité va dépendre de la qualité du mélange des différentes « grammaires » professionnelles désormais mobilisées, ainsi que de la profondeur de l'acculturation des journalistes aux facettes marketing de leur activité. Au sein de chaque rédaction, l'éventail des compétences mobilisées va devoir s'ouvrir comme jamais (Desbordes, 2018, p. 124). A chaque compétence sont associés une « grammaire », des techniques et des savoir-faire. Les rédacteurs vont devoir interagir avec des spécialistes de design graphique (visualisation des données), des *community managers*, des *data miners*, des *data analysts*, des spécialistes d'ergonomie cognitive, ... et partager avec ces différents professionnels des connaissances, des diagnostics et des outils. L'adaptabilité sera donc affaire de polyvalence, de « langue commune » et de croisement de plus en plus marqué entre des univers techniques jadis dissociés. Bien au-delà de la simple problématique RH (gestion des compétences et des qualifications), cette nécessaire « hybridation » soulève la question de la formation au journalisme (qu'elle soit initiale ou continue). En France où les canons de la profession demeurent encore polarisés autour de référentiels foncièrement littéraires, une telle mutation ne va pas de soi et pose un défi quasi-culturel.

Autre défi de taille, celui de l'acculturation des journalistes aux facettes marketing de leur activité. Comme l'a bien montré Pauline Amiel dans son étude récente des journalistes « localiers » de la presse quotidienne régionale française, cette acculturation se heurte à un conflit d'identité. A l'heure de la dématérialisation de la presse locale et de la mutation de son modèle économique, ces « localiers » sont tiraillés entre une identité professionnelle traditionnelle (« fermée », c'est-à-dire bâtie en opposition avec les canons de la communication) et une identité moderne plus ouverte, plus mouvante, assise sur une « intégration du vocabulaire et des enjeux » (Amiel, 2017, p. 221) du marketing. Ces professionnels vivent souvent douloureusement ce qu'ils assimilent à une « dilution [...] de l'identité énonciative » (*id.*, p. 302) des médias d'information. Ce ressenti collectif peut à l'évidence freiner l'adaptation au nouveau contexte précédemment évoqué.

La reconfiguration organisationnelle ne s'arrête bien sûr pas aux portes de la rédaction ; elle impacte la firme médiatique *lato sensu.* Les médias, en tant qu'organisations économiques, seront soumis au décloisonnement interne. En mai 2014, le bilan « *Innovation Report* » qui tente de souligner les points faibles de la stratégie numérique du *New York Times* pointe, entre autres choses, la nécessité d'une coopération, ou d'un maillage plus étroit, entre la rédaction du journal, son service marketing et le service « expérience-client » (lequel rassemble, à l'époque, 30 designers numériques, 30 analystes de données, 120 chargés du développement-produit et 445 informaticiens ingénieurs ! (Fradin, 2016, p. 85)). En France, les expériences de « labs », mises en place récemment au sein de certains médias (*Le Parisien*, par exemple), participent de cette même logique de décloisonnement. Une dimension « *bottom-up* » est souvent mise en exergue, laquelle – en permettant une vraie implication des journalistes dans la recherche de solutions nouvelles – facilite l'acceptation, par ces derniers, du débordement des frontières traditionnelles de leur métier. L'enjeu de cette acculturation n'est pas mince : il s'agit de catalyser, au sein des rédactions, un processus permanent d'innovations (nouveaux formats éditoriaux, nouvelles interfaces, ...), seules susceptibles de concilier *in fine* les attentes labiles des consommateurs, par ailleurs « citoyens-internautes », avec les savoir-faire et les ambitions légitimes des professionnels de l'information ; comme si la révolution des « *data* » permettait aux médias de densifier les liens, jadis distendus, avec leurs publics.

Conclusion

Dès lors que l'on s'interroge sur les impacts du traitement algorithmique des données massives sur l'univers du journalisme professionnel, force est de constater la polysémie qui auréole la notion de *big data*. De quelles données parle-t-on ? De celles qui traduisent un nouveau rapport au terrain d'investigation (les sources), ou bien de celles qui révèlent les caractéristiques comportementales des destinataires de l'information ? ... Ce distinguo est important, comme nous avons tenté d'ailleurs de l'établir. Cette polysémie est d'autant plus significative qu'elle ouvre la voie, du fait des différentes acceptions qu'elle véhicule, à des visions très contrastées de l'avenir du journalisme professionnel. Une vision ultra-techniciste, adossée à la « science » des *big data*, alimenterait pour sûr la prophétie de la disparition du métier ; à l'aune d'une telle vision, une production automatisée d'information, tirée en temps réel par la demande, se substituerait avantageusement à la médiation humaine constitutive de l'exercice de la profession. A cette vision peut être opposée la vision plus nuancée que nous formulons ici. Loin de « tuer » le métier, la déferlante des *big data* va sans doute, et de façon presque paradoxale, contribuer à sa mutation en renforçant deux de ses caractéristiques fondamentales, à savoir le tropisme en faveur du terrain ainsi que l'approfondissement des liens avec le public.

Références bibliographiques

Amiel, Pauline (2017), *L'identité professionnelle des localiers à l'heure des mutations économiques et de la dématérialisation de la presse locale*, Thèse de Doctorat de l'Université de Toulouse-Paul Sabatier, Toulouse, 24 novembre.

Anderson Chris (2008), "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", *Wired*, June 23rd (consulté le 13 février 2018, http://www.wired.com/2008/06/pb-theory/).

Anderson Christopher W. (2015), "Between the unique and the pattern. Historical tensions in our understanding of quantitative journalism", *Digital journalism*, n°3, p. 349-363.

Bassoni Marc, Joux Alexandre (2014), *Introduction à l'économie des médias*, Paris : Armand Colin (Cursus).

Boyd Danah, Crawford Kate (2012) "Critical questions for big data", *Information, Communication & Society*, 15/5, p. 662-679.

Cardon Dominique (2013), « Dans l'esprit du PageRank. Une enquête sur l'algorithme de Google », *Réseaux* 2013/1, n° 177 : 63-95

Cardon Dominique (2015), *A quoi rêvent les algorithmes ? Nos vies à l'heure des big data*, Paris : Seuil (La République des idées).

Carlson Matt (2015), "The Robotic Reporter", *Digital Journalism*, n°3, p. 416-431.

Coddington Mark (2015), "Clarifying journalism's Quantitative Turn", *Digital Journalism*, n°3, p. 331-348.

Croissant Valérie, Touboul Annelise (2013), « Le *datajournalism* par ses acteurs, autodéfinition d'une pratique émergente en France », *Recherches en communication*, 40, p. 133-149.

Dagiral Eric, Parasie Sylvain (2011), « Portrait du journaliste en programmeur : l'émergence d'une figure du journaliste "hacker" », *Les Cahiers du journalisme*, n° 22/23, p. 144-154.

Daniel Anna, Flew Terry (2010), "The Guardian Reportage of the UK MP Expenses Scandal: a Case Study of Computational Journalism", *Communications Policy & Research Forum*, Sydney, November 15-16th.

Desbordes Damien (2018), *Les robots vont-ils remplacer les journalistes ?*, Paris : Plein Jour.

Fradin Selma (2016), *Les nouveaux business models des médias. Les trois piliers de la transformation*, Fyp (Entreprendre).

Gayant Jean-Pascal (2015), « Le big data profite-t-il vraiment au consommateur ? », Le Monde, 17 janvier.

Gillespie Tarleton (2014), "The relevance of Algorithms" (p. 167-194), in *Media Technologies: Essays on Communication, Materiality and Society* (Gillespie and alii dir.), Cambridge (MA): MIT Press.

Guibert Gérôme, Rebillard Franck, Rochelandet Fabrice (2016), *Médias, culture et numérique. Approches socioéconomiques*, Paris : Armand Colin (Cursus).

Gyldensted Cathrine (2015), *From Mirrors to Movers: Five Elements of Positive Psychology in Constructive Journalism*, Seattle : CreateSpace Independent Publishing Platform.

IAB (Europe) (2017), *European programmatic market sizing 2016*, september (consulté le 13 février 2018, https://www.iabeurope.eu/research-thought-leadership/programmatic/iab-europe-report-european-programmatic-market-sizing-2016/).

Kayser-Bril Nicolas, Gray Jonathan, Bounegru Liliana, Chambers Lucy (dir.) (2013), *Guide du datajournalisme : collecter, analyser et visualiser les données*, Paris : Eyrolles.

Kessous Emmanuel (2011), « L'économie de l'attention et le marketing des traces », Communication au colloque ACFAS *Web social, communautés virtuelles et consommation*, UQAM, Sherbrooke, 11 mai.

Lewis Seth C., Westlund Oscar (2015), "Big data and journalism. Epistemology, expertise, economics and ethics", *Digital journalism*, 3, p. 447-466.

McManus John (1994), *Market-Driven Journalism: Let the Citizen Beware?*, London : Sage Publications.

Meyer Philip (1973), *Precision journalism: a reporter's introduction to social science methods, Bloomington*, Indiana University Press.

Ouakrat Alan (2012), « Le ciblage comportemental, une perte du contrôle des éditeurs sur les données de l'audience », *Tic&sociétés* 6(1) : 33-55

Parasie Sylvain, Dagiral Eric (2013), « Des journalistes enfin libérés de leurs sources ? Promesses et "réalités du journalisme de données" », *Sur le journalisme* [en ligne], Vol. 2, n°1, p. 52-63.

Ruellan Denis (1992), « Le professionnalisme du flou », *Réseaux*, Vol. 10, n° 51, p. 25-37.

Scherer Eric (2015) *Logiciels, algorithmes, robots : journalisme automatique*, méta-media#8, Paris, France Télévisions

Schudson Michael (2001), "The objectivity norm in American journalism", *Journalism*, 2/2, p. 149-170.

Welsh Ben, Smith Doug (2009), "Highest crime rate in L.A.? No, just an LAPD map glitch", *latimes.com*, April 5th (consulté le 8 février 2018, http://www.latimes.com/local/la-me-geocoding-errors5-2009apr05-story.html).

Les réinventions de la démocratie à l'aune de l'ouverture des données : du discours de la participation aux contraintes de la gouvernance

Reinventions of democracy in the light of open data:
from the discourse of participation to the constraints of governance

Reinvenciones de la democracia a la luz de los datos abiertos:
del discurso de la participación a las limitaciones de la gobernanza

Article inédit, mis en ligne le 15 novembre 2018.

Anne Lehmans

Anne Lehmans (anne.lehmans@u-bordeaux.fr) est maître de conférences en sciences de l'information et de la communication à l'Université de Bordeaux (ESPE d'Aquitaine). Elle appartient à l'équipe de recherche RUDII (Représentations, usages, développements et ingénieries de l'information) du groupe Cognitique à l'IMS (Intégration matériaux-système, CNRS-UMR5218).

Plan de l'article

Résumé

Une recherche sur les politiques et les pratiques de médiation, de valorisation et d'éducation autour des données ouvertes (*open data*), révèle que celles-ci peinent à trouver leur place dans l'espace public mais sont pourtant porteuses d'une interrogation fondamentale sur le rôle des données dans le fonctionnement de la démocratie contemporaine. L'espace public des données ouvertes est déterminé dans des lieux de négociation et selon des arrangements peu lisibles. Les dispositifs qui permettent l'accès à ces données répondent à une logique de gouvernance administrative et économique plus que démocratique. Ainsi, pour que les données ouvertes participent d'une réinvention de la démocratie dans le contexte du *big data*, les pratiques informationnelles et cognitives qui permettent les usages des données nécessitent des formes d'acculturation qui sont complexes.

Mots clés

Open data, gouvernance des données, médiation, culture de l'information.

Abstract

Research on policies and practices of mediation, valuation and education on open data reveals that the opening of data triggers new questions that carry a fundamental question about the place of data in the functioning of contemporary democracy. The public open data space is determined in negotiating locations and in unclear arrangements. The devices that allow access to these data respond more to a logic of administrative and economic governance than to democratic principles. Thus, for open data to participate in a form of reinvention of democracy in the context of big data, the information and cognitive practices that allow the use of data require complex acculturation actions and the development of data literacy.

Keywords

Open data, data governance, mediation, information literacy.

Resumen

La investigación sobre políticas y prácticas de mediación, evaluación y educación en torno a los datos abiertos revela que es difícil encontrar un lugar en el dominio público de la democracia contemporánea. El espacio público de datos abiertos se determina en las ubicaciones de negociación y en los acuerdos poco claros. Los dispositivos que permiten el acceso a estos datos responden a una lógica de gobierno administrativo y económico que es más que democrática. Por lo tanto, para el propósito de la reinvención de la democracia en el contexto del big data, la información y las prácticas cognitivas que permiten el uso de datos requieren acciones de aculturación que son complejas.

Palabras clave

Datos abiertos, gobierno de datos, mediación, cultura de la información.

Introduction

Parmi la masse des données qui alimentent le *big data*, les données ouvertes, *open data*, occupent une place particulière. Elles désignent les données collectées par les organismes publics ou privés chargés d'un service public et mises à disposition en format numérique sur des plateformes nationales ou locales permettant leur libre accès et leur réutilisation par les citoyens ou les organisations. Plusieurs types de données peuvent faire l'objet d'une ouverture et d'une mise à disposition pour la société (Liquète, Gilliard, 2017). Différents territoires sont concernés, dans la logique affirmée par la mission Etalab en 2011, confirmée par la création de la fonction d'administrateur général des données (décret n°2014-1050 du 16 septembre 2014), la participation

de la France au Partenariat pour un Gouvernement Ouvert en 2014 et la loi pour une République numérique en 2016. Cette dernière a ajouté la notion de "données d'intérêt général" qui peuvent être diffusées par les acteurs privés en lien avec des données publiques, comme ceux qui perçoivent des subventions publiques. Les données provenant de travaux de recherches scientifiques subventionnés par l'Etat en font partie. Un livre blanc sur les données ouvertes (Meszaros, Samath, Guerin-Hamdi, Faure, 2015) indique que les enjeux de ces dernières concernent les sphères socio-économiques, scientifiques et culturelles pour permettre une meilleure capacité intégrative de l'individu et du groupe social, des possibilités d'innovation, des potentialités de réaction et d'adaptation à l'évolution de l'organisation individuelle ou collective. Le portail de modernisation de l'action publique voit dans *l'open data* un levier du changement de l'administration, affirmant que « *l'*open data *nourrit la participation citoyenne et vice-versa, confère de nouveaux moyens d'agir et stimule la démocratie* »[1].

Mises à disposition du public, les données ouvertes découvrent un champ potentiel de connaissances et d'utilisations multiples, dans une perspective technologique de « *smart city* », mais aussi de démocratie participative, à tous ceux qui sont en capacité de les appréhender et d'en faire usage. L'ouverture des données répond à un enjeu politique de transparence de l'action publique à travers une conception très large du droit à l'information des citoyens initiée en France par la loi du 17 juillet 1978 sur la liberté d'accès aux documents administratifs et de la réutilisation des informations publiques. Ce droit concerne le citoyen non seulement du point de vue de la protection de ses données personnelles (dans la logique confirmée par le Règlement général sur la protection des données qui entre en vigueur en mai 2018), mais également de sa possibilité de contrôler l'action publique, voire d'y participer directement. Ainsi, Sarah Labelle et Jean-Baptiste Le Corf (2012) ont montré que les politiques d'ouverture des données s'accompagnent de discours qui visent le politique à travers la qualité des processus démocratiques, l'administratif à travers une amélioration des relations entre administration et administrés, et le technique à travers le soutien à l'innovation. Dans la perspective de valoriser l'ensemble de ces objectifs, les enjeux de l'ouverture sont aussi communicationnels (Capelle, Lehmans, 2016). L'expression *Open data* est initialement liée à la recherche publique environnementale. Le concept a été créé pour répondre à un triple besoin : partager les données pour favoriser une réflexion globale, baisser le coût de ce partage, et favoriser la participation (Goncalves, Rufat, 2016). Concrètement, sur les portails d'*open data*, des jeux de données peuvent être téléchargés librement en vue d'un traitement qui permet, par exemple, de vérifier des hypothèses, de créer des applications, ou de représenter visuellement des informations grâce à des outils graphiques. Ils constituent pour le citoyen à la fois une fenêtre sur l'activité des services publics, une manne informationnelle alimentant le débat public, et un espace de participation tant par la réutilisation des données que par la contribution à leur collecte. Selon les jeux de données, la manière dont ils sont présentés, et l'exploitation qui est visée, les opérations techniques et cognitives nécessaires peuvent s'avérer complexes.

Si la question des données ouvertes évolue rapidement dans les agendas politiques, du local à l'international, la réalité des pratiques en France reste hétérogène (Goeta, 2016). L'affichage d'une politique d'ouverture cache les contraintes fortes qui pèsent sur la gouvernance des données. Il n'est pas exempt de discours idéologiques, comme le montre Etienne Damome (2018) à propos du cas africain, dans lequel le développement et la démocratisation sont au cœur d'un « fantasme techno-déterministe » qui accompagne la politique d'ouverture des données. Plus profondément, cette question interroge les fondements mêmes de la démocratie, quand la gestion des données risque d'envahir ou de remplacer l'espace public et d'en effacer les dimensions proprement politiques du débat, d'en faire oublier la nature contraignante. Une recherche sur les politiques et les pratiques de

[1] http://www.modernisation.gouv.fr/ladministration-change-avec-le-numerique/par-louverture-des-donnees-dans-les-administrations/open-data-nourrit-la-participation-citoyenne

médiation, de valorisation et d'éducation autour des données ouvertes a été menée par notre équipe dans la région Nouvelle Aquitaine[2]. L'enquête a porté sur plusieurs entités productrices et utilisatrices de données ouvertes dans l'académie de Bordeaux. Son objectif était d'analyser, d'une part, les politiques d'ouverture, de médiation et de médiatisation des données à l'échelon territorial, d'autre part, la façon dont les enseignants recourent effectivement à ces données ouvertes dans un objectif pédagogique. Elle révèle que l'ouverture des données déclenche des questions nouvelles qui peinent à trouver leur place dans l'espace public mais qui sont pourtant porteuses d'une interrogation fondamentale sur le fonctionnement de la démocratie contemporaine, à l'heure où l'extrême quantité, la variété et la vitesse de circulation des données qui caractérisent le *big data* semblent remettre en cause les processus traditionnels de prise de décision.

Alors que les pratiques journalistiques se déplacent sur la manipulation de données (*fact checking*, lutte contre les *fake news*) et de la *data visualisation*, que l'école s'ouvre de plus en plus aux techniques de gestion des données pour construire des connaissances, que les discours politiques s'appuient largement sur l'objectivité et la transparence apparentes des données et des chiffres, un temps de réflexion sur le sens et les usages des données ouvertes est indispensable. Quelle place sont-elles susceptibles d'occuper dans l'espace public entendu comme un espace de communication politique, du point de vue organisationnel, des représentations et des pratiques ? L'espace public de la donnée ouverte est un espace symbolique de discours articulé aux dispositifs socio-techniques de mise à disposition et d'usage des données (Lehmans, 2018). L'ouverture des données est associée dans les discours à la gouvernance ouverte, voire au gouvernement ou à la démocratie ouverte. Techniquement, les données ouvertes se présentent sous forme de bases contenant des jeux de données, constituées et mises à disposition dans le cadre d'une collecte réalisée par différents acteurs aux stratégies complémentaires, concurrentes ou divergentes. Elles n'existent pas tant qu'elles ne sont pas mises en visibilité, et socialement appropriées. La donnée est produite dans un contexte et offerte dans une perspective d'usage projeté. Il est donc important d'identifier les conditions de cette production d'une part, les projets d'usages d'autre part, les conditions de réalisation de ces projets enfin, pour les situer dans la perspective politique d'une évolution des mécanismes démocratiques. Dans un premier temps, il faut souligner que l'espace public des données est déterminé dans des lieux de négociation et selon des arrangements qui doivent être décryptés. Ensuite, l'analyse des dispositifs qui permettent l'accès aux données montre qu'ils répondent à une logique de formatage organisationnel et technique à considérer en fonction des préoccupations de participation. Enfin, les pratiques informationnelles et cognitives qui permettent les usages des données nécessitent des formes d'acculturation qui sont complexes et peu négociables, et pourtant conditions *sine qua non* de l'occupation maîtrisée d'un espace public de la donnée transformable en lieu de décision et de participation politique démocratique.

Gouvernance ouverte et lisibilité de l'espace public de la donnée

Les données ouvertes relèvent d'une politique volontariste de partage et de collaboration dans la production et les usages de l'information à partir de données mises à disposition. Du point de vue des politiques publiques, elles reposent sur la mise en œuvre de principes affichés de transparence, de participation citoyenne et de modernisation de l'action publique par la collaboration entre les institutions et les citoyens (El Hachani, 2015 : 5). Le premier principe de transparence se heurte aux contraintes organisationnelles des entités productrices de données.

. . . . . .

[2] L'équipe RUDII (Représentations, usages, développements et ingénierie de l'information) de l'IMS (Intégration matériau-système, UMR5218), qui a mené le projet de recherche dans l'académie de Bordeaux. Le carnet de recherche Data-cultures est disponible : http://dcultures.hypotheses.org/

Des stratégies politiques de mise en visibilité des données

Avant la loi pour une République numérique, qui pose le principe de l'ouverture des données publiques par défaut, la décision d'ouverture et de diffusion des données est souvent politique et parfois liée à une concurrence entre collectivités sur un territoire ou entre services dans une collectivité. Elle révèle des stratégies de communication politique autour de la transparence et de l'innovation qui peuvent être mises en avant par des collectivités de taille très modeste comme la commune de Brocas Les Forges dans les Landes (800 habitants), qui affiche une politique volontariste d'ouverture des données depuis 2011 et un site avec des propositions de réutilisations des données. Ainsi, l'ouverture des données s'inscrit dans une rhétorique de la visibilité, de l'innovation et de la participation, parfois autour de la thématique plus précise de l'économie sociale et solidaire. La conséquence de ces stratégies est que les données sont souvent orientées vers les informations géographiques et liées à des territoires, et que chaque territoire développe ses propres pratiques d'ouverture. Des phénomènes d'asymétries d'information entre acteurs et territoires (Chartron, Broudoux, 2015) risquent d'entrer en contradiction avec la logique d'ouverture des données qui, pour être efficace à grande échelle et produire des effets durables, ne peut pas être cloisonnée. En outre, les logiques communicationnelles d'ouverture ne sont pas toujours congruentes avec les logiques organisationnelles.

Le projet de recherche a permis d'enquêter entre 2016 et 2017 sur le terrain de collectivités locales productrices de données auprès d'acteurs en Nouvelle Aquitaine, des médiateurs et des utilisateurs finaux des jeux de données ouvertes. Du côté des producteurs de données, personnels de collectivités, médiateurs, élus, des observations participatives de séances de travail sous forme d'ateliers et de focus groupes ont été réalisées, ainsi que des entretiens semi-directifs auprès de six agents, afin d'identifier et de caractériser le discours des acteurs de terrain pour comprendre les objectifs qu'ils se fixent, et leur représentation de la chaîne de production et de diffusion des données dans la sphère publique actuelle. Les entretiens permettent de comparer les représentations d'acteurs entre des secteurs d'activités ou des échelles géographiques différents, de comprendre l'articulation de la chaîne de production et de diffusion des données, et de saisir l'évolution de leurs pratiques déclarées sur leurs terrains. Du côté des utilisateurs dans un objectif d'éducation, l'équipe de recherche s'est intéressée aux usages des enseignants. Elle l'a fait à partir d'entretiens menés auprès de quinze d'entre eux, intervenant à l'université, dans la formation des professeurs, ou au rectorat dans la formation continue, et d'observations de séances de formation d'étudiants ou d'enseignants. Tous les enseignants interrogés étaient des utilisateurs potentiels ou avérés de données ouvertes dans leurs pratiques pédagogiques.

Plusieurs plateformes d'*open data* ont vu le jour au sein des collectivités publiques pour présenter et mettre à disposition leurs jeux de données. En France, Etalab a mis en place la plateforme data.gouv.fr, tandis que des collectivités territoriales ont créé leurs propres plateformes de données locales et que des entreprises comme la SNCF, mais aussi Uber ou AirBnB, communiquent sur l'ouverture de (certaines de) leurs données. Dans la Région Nouvelle Aquitaine, les efforts se sont multipliés pour mettre à disposition des données dans le domaine du tourisme (SIRTAQUI), de l'environnement et des données géographiques (PIGMA, Datalocale), par exemple. Ces plateformes numériques reposent sur des stratégies de mise en visibilité des données (Mabi, 2015) par une grande variété d'acteurs qui sont entrés dans des processus d'ouverture pour des raisons diverses : stratégiques et économiques, lorsque que le calcul coût-bénéfice penche en faveur de l'ouverture pour des plateformes d'information géographique par exemple, politiques, lorsque des élus s'emparent de la question des données ouvertes dans une perspective militante ou d'image, normatives, lorsque la loi impose l'ouverture des données. Dans les collectivités territoriales, la direction numérique, à partir d'une demande politique ou de sa propre initiative, donne souvent l'impulsion aux propositions qui sont validées suivant un schéma administratif puis par les élus qui

interviennent en fin de processus. Ces derniers n'ont pas toujours une compréhension globale et précise des enjeux de l'ouverture des données.

Des contraintes organisationnelles de faible visibilité des données

Malgré la législation récente, les administrations sont encore loin de mettre à disposition toutes les données qu'elles gèrent. Les agents des collectivités territoriales témoignent d'une grande diversité des pratiques entre les services, qui ne sont pas tous, du point de vue technique, politique, ou managérial, prêts à considérer cette gestion dans un écosystème ouvert. L'ouverture des données implique la transformation des structures et des modes de gouvernance des organisations ainsi que des compétences des agents. Dans le rapport remis par le Secrétariat général pour la modernisation de l'action publique en 2015, la gouvernance de la donnée désigne « *l'ensemble de principes et de pratiques qui visent à assurer la meilleure exploitation du potentiel des données* » (p. 48). Les principes relèvent d'une logique de transversalité dans l'organisation, les pratiques d'une dynamique de communication. Le rapport pointe les limitations dans le potentiel des usages des données, liées à la méconnaissance, à l'imperméabilité de la culture administrative par rapport aux dynamiques de coopération et de participation. Il préconise une évolution des systèmes d'information et un décloisonnement des administrations. La gouvernance « *associe négociations, prises de décision et émergence de consensus* » (Schafer et Le Crosnier, 2011). La gouvernance ouverte lie les principes de transparence et de participation dans une relation complexe intégrant visibilité de l'action publique, développement des moyens d'expression, voire d'action, des citoyens et "comportements organisationnels" ou logiques bureaucratiques qui peuvent constituer des entraves à l'ouverture (Pasquier, Villeneuve, 2007).

L'ouverture des données dans une organisation questionne ainsi ses modes de fonctionnement interne, les compétences professionnelles ou « métiers » dans et entre services, la façon d'intégrer le public susceptible d'utiliser les données dans la prise de décision, et, de façon générale, l'écosystème informationnel et les interactions entre les services et les acteurs. Les méthodes de collecte, de traitement, de conservation et de communication des données sont propres à chaque service, qui fonctionne avec la direction des systèmes d'information, mais pas de façon transversale et partagée avec les autres services. Les problématiques d'accessibilité et d'utilisabilité perturbent les fonctionnements routiniers de services cloisonnés dans leur gestion des données, de l'information et des documents, et mettent en relief sinon des dysfonctionnements, du moins des résistances au changement qui les placent en porte-à-faux avec les services chargés de la communication, d'une part, les politiques, d'autre part, porteurs d'un discours d'ouverture et de transparence. Le dossier de l'ouverture des données est impulsé par un service en particulier, souvent la direction des systèmes d'information qui en fait un levier de modernisation des pratiques, mais parfois aussi les services nouveaux créés autour des questions d'usages, portés par des agents dont les compétences sortent du seul domaine de l'informatique et sont en lien avec la gestion de l'information d'une part, l'intelligence économique d'autre part, l'animation des données enfin. Ainsi, pour l'un des acteurs interrogés, « *dans une organisation hyper-hiérarchique, soit on sensibilise les cadres, soit on fonctionne en mode projet. On essaie de ne pas faire de l'open data une fin en soi, sinon on recrée un silo.* »

De nombreux métiers nouveaux se sont développés autour des stratégies d'ouverture des données, qui n'ont pas encore totalement acquis leur identité dans la langue française puisque l'on continue de parler de « *data scientist* », « *data analyst* », « *data manager* », « *chief data officer* » ou « *data protection officer* », cette dernière fonction bientôt rendue obligatoire par l'application du Règlement Général sur la Protection des Données européen. La question de la donnée dans le *big data* est un enjeu de pouvoir central dans le fonctionnement des organisations aujourd'hui, et la problématique des données ouvertes renforce encore la position des professionnels de l'informatique, tout en nécessitant des formes de coopération et de communication avec les professionnels de l'information

qui ont une vision des enjeux économiques et sociaux. La culture de la donnée n'est pas strictement technique mais aussi analytique, liée à la capacité d'identifier les enjeux et les données susceptibles de créer de l'information qui a de la valeur. La gouvernance reste élitiste dans ses principes de fonctionnement, même si ceux qui portent les projets d'ouverture et de mise à disposition des données se vivent souvent comme des missionnaires. Ainsi, dans une collectivité, c'est un « conseil de sachants » qui donne l'impulsion dans la prise de décisions, parce que ceux qui savent connaissent les contraintes et sont conscients des enjeux. Les enjeux politiques de l'ouverture des données sont fortement corrélés à des questions économiques dans le cadre global du *big data*.

Formatage et utilisabilité de l'espace public de la donnée

Au-delà de la prise en considération des risques et de contraintes techniques et économiques fortes, la gouvernance des données ouvertes interroge la possibilité d'une participation citoyenne à la gestion du cycle de vie de l'information. Mais cette question, au cœur des principes de gouvernance de l'information dans les entreprises, n'apparaît pas de façon évidente dans les discours des acteurs et dans les normes concernant les données ouvertes. Elle appelle des formes d'explicitation des décisions et des procédures et de responsabilisation aussi bien des acteurs de l'ouverture que des usagers des données. Elle est complexe, en rupture par rapport à la culture administrative française centralisée, et croise des questions qui relèvent de l'expertise technologique. Des types de légitimité contradictoires sont en jeu, l'une axée sur la démocratie et le dialogue autour de la thématique de la participation, l'autre sur l'expertise technique et l'efficacité autour de la thématique de la modernisation et de l'innovation. C'est clairement l'entrée des usagers et la question de l'utilisabilité des données dans l'écosystème qui imposent la nécessité de repenser les formats d'interopérabilité sur le plan technique.

Repenser les formats

Avant la mise en place d'une dynamique d'ouverture, les données étaient collectées et les flux traités et régulés dans les systèmes internes d'information sans que les questions de l'ouverture, de la publication et finalement de la communication vers l'extérieur fussent prises en compte. L'ouverture des données réinterroge le cycle de vie de la donnée en contraignant les organisations et les services à prendre en considération, dès le moment de la collecte, la possibilité de réutilisation, sans que celle-ci soit clairement définie à l'avance. En effet, la mise à disposition des données n'implique pas automatiquement leur usage. Le volume, la diversité et l'hétérogénéité des données nécessitent un pilotage dans le traitement de l'infrastructure technique et logicielle ainsi que dans la description et la mise en forme des données en vue de leur publication, dans une logique de formatage comme condition du partage. Le pilotage doit intégrer la valeur des données au regard des usages projetés, de la transformation de la donnée en information, dans une démarche de qualification qui rompt avec les routines organisationnelles et qui nécessite de la part des agents une expertise dans la représentation des bases de données à l'aune de leurs usages externes. L'utilisabilité des données doit ainsi être considérée, définie par la norme comme « *le degré selon lequel un produit peut être utilisé, par des utilisateurs identifiés, pour atteindre des buts définis avec efficacité, efficience et satisfaction, dans un contexte d'utilisation spécifié* » (ISO 9241-11:2018). Le format informatique, la licence, l'organisation de la base de données, conditionnent l'utilisabilité et la réutilisabilité des données ouvertes et les installent dans une logique de normalisation.

Les collectivités ont commencé à mettre en place des formes de réorganisation des procédures et des métiers quand elles se sont lancées volontairement dans des politiques d'ouverture. Elles sont désormais contraintes de le faire avec la mise en application du Règlement général pour la protection des données, qui impose le principe de responsabilité par rapport aux usages. Outre la gestion de la protection des données personnelles, une dimension incontournable dans cette réorganisation est

celle de la normalisation qui porte des enjeux juridiques et techniques mais aussi fonctionnels. Les normes et les formats assurent le lien entre collecte, accès et usages tout au long du cycle de vie de la donnée autour duquel les compétences des agents se réforment. Dans un contexte de territorialisation des données ouvertes et de coexistence entre des pratiques de gestion des données diverses, l'absence de normalisation freine l'interopérabilité, la garantie de la qualité et la possibilité de diffuser des jeux de données universellement utilisables.

Standardisation et usages de l'information

La standardisation peut constituer un levier de gouvernance des données, quand elle devient un objet partagé dans et entre les services. Dans le domaine de l'information géographique, la directive INSPIRE de la Direction générale de l'environnement de la Commission européenne datant de 2007, « *vise à établir en Europe une infrastructure de données géographiques pour assurer l'interopérabilité entre bases de données et faciliter la diffusion, la disponibilité, l'utilisation et la réutilisation de l'information géographique en Europe* » (Conseil national de l'information géographique). Dans ce cas, c'est au niveau politique et à l'échelon européen que des standards ont été imposés afin d'obtenir une information structurée, mise à jour et partageable, qui a permis le développement d'outils comme le Géocatalogue et le Géoportail et à des plateformes comme PIGMA de fonctionner. L'application de cette directive oblige, dans les organisations chargées de l'information géographique, les thématiciens et les informaticiens à adopter des normes, des langages, des pratiques communs visant l'interopérabilité des données. Une véritable culture technique et normative s'est développée sur la base de cette directive pourtant très complexe, et a donné aux services d'information géographique une longueur d'avance sur les collectivités peu sensibilisées aux nécessités de standardisation et à des outils techniquement complexes. Cette standardisation est la condition de la mutualisation et les médiateurs de données la mettent au cœur de leur activité en prenant en charge la formation des services producteurs aux standards et à la structuration des données.

Dans une logique de partage de la donnée, les formats conditionnent l'exploitabilité des données. L'utilisation d'un format ouvert (de type .csv) pour mettre à disposition les jeux de données est essentielle. Les licences qui déterminent les conditions juridiques de réutilisation des données sont également fondamentales. Elles devraient privilégier le partage à l'identique, mais celui-ci est contraire aux intérêts commerciaux des entreprises et un frein à l'utilisation des données ouvertes. La plupart des collectivités territoriales et l'association Open Data France ont donc renoncé à imposer des licences trop strictes (OdBL) imposant un partage à l'identique et fait le choix pragmatique de permettre aux entreprises de privatiser les services produits à partir des données ouvertes (licence OL créée par Etalab). Enfin, la qualité des métadonnées est très importante pour permettre la contextualisation des jeux de données. Charlotte Maday (2015 : 161) souligne ainsi la proximité de la gestion des données ouvertes avec le *record management*.

L'utilisation réelle et contextualisée des données requiert le respect de normes et de standards dont les exigences techniques ne sont pas toujours aisées à comprendre, notamment des élus, et qui conditionnent pourtant l'efficacité des politiques d'ouverture. Mais celles-ci, au niveau territorial, doivent avant tout rencontrer des besoins et des usages, une culture partagée de la donnée.

Cultures de la donnée, démocratisation de l'espace public de la donnée et communs de la connaissance

L'ouverture reste une utopie (Goeta, 2016) si les conditions organisationnelles et culturelles de leur usage ne sont pas réunies. Même si les données sont accessibles, sur les plans technique et juridique, le « public » visé par l'ouverture doit connaître leur existence d'une part, être capable de les utiliser d'autre part. Les responsables des services chargés de l'ouverture des données le savent, le public

ciblé reste nébuleux. Les entreprises, ou les services publics qui ont besoin de données, sont actuellement les principaux bénéficiaires de l'ouverture des données. Les conditions d'une véritable réinvention ou rénovation de la démocratie à l'aune de la participation de tous aux usages des données, dans le cadre de communs informationnels (Peugeot, 2016), restent à construire, par des stratégies de médiation et d'éducation. Ce qui pourrait constituer la part démocratique du *big data* à travers la création d'un espace commun des données n'est pas encore d'actualité. Cet objectif repose sur deux exigences de base qui sont prises en charge par les organisations : la première est que les données soient visibles et lisibles par tous grâce à des bases de données permettant de les rechercher aisément ; la seconde est qu'elles soient suffisamment accessibles, pour être utilisées par des acteurs qui n'ont pas participé à leur processus de fabrication. Une troisième dimension reste essentielle, mais elle est encore peu envisagée : le développement d'une littératie des données ouvertes à travers des stratégies de médiation et d'éducation. C'est cette dimension qui a été analysée dans le projet de recherche.

Les stratégies de médiation

Les objectifs de soutien à l'innovation avec l'*open data* énoncés dans la loi Pour une République Numérique (2016), et l'ouverture d'une manne informationnelle en direction de la société civile, peuvent favoriser de nouveaux usages à partir des données. Mais l'innovation concerne essentiellement, pour l'heure, les moyennes ou grandes entreprises qui disposent de ressources humaines et techniques suffisantes pour valoriser les données, à la différence des petites entreprises et des citoyens. L'usage des données, dans le monde du *big data* et de l'information, est complexe et coûteux, tant il suppose l'intervention de spécialistes. En outre, les données ne sont utilisables que dans le cadre d'un travail complexe de traitement et de construction de l'information à la base des connaissances. Il apparaît alors essentiel de développer des offres de services et de formations qui permettent aux citoyens et aux entreprises d'utiliser les données ouvertes telles qu'elles sont produites et mises à disposition, en vue de répondre à des besoins d'information pour agir, dans une logique d'innovation et d'économie sociale et solidaire.

Quels que soient les moyens de médiation mis en place sur les plateformes, la manipulation des données par traitement informatique à l'aide de machines suppose la maîtrise de compétences de la part des utilisateurs. Sur la plateforme data.gouv.fr, les "meilleures" réutilisations sont rendues visibles et mises en avant sur la page d'accueil, ce qui valorise les usages possibles et invite les contributeurs potentiels à la créativité. Des explications sur les différentes méthodes, moyens et outils techniques accompagnent aussi les usagers qui souhaitent se lancer en tant que contributeurs, mais de façon très sommaire. Dans les collectivités territoriales, des actions de médiation sont mises en place en interne pour diffuser la culture des données ouvertes dans et entre les services, souvent avec l'aide de médiateurs extérieurs comme la Fondation internet nouvelle génération, qui organise des « Infolabs ». En direction de l'extérieur et des publics, outre la médiation numérique *via* l'organisation et la communication autour des plateformes, ce sont plutôt les « animations » qui sont privilégiées ; cartoparties, ateliers de réflexion et de data-visualisation, concours, appels à projets permettent de mettre en scène les données ouvertes en direction de publics précisément ciblés et dont la détermination est liée aux missions et au ciblage politique de la collectivité.

Cependant, les seules visibilités en ligne et animations des données ouvertes ne permettent pas aujourd'hui de répondre aux promesses de l'*open data*, en matière de réutilisation des données par tous, d'intelligence économique et plus généralement d'innovation et de participation. Le développement des compétences liées au numérique et plus précisément aux données est indispensable, comme le rappelle la création par un décret du 3 novembre 2017 du haut-commissaire à la transformation des compétences, auprès du ministre du travail. Cette nécessité était affirmée par un avis du Conseil économique, social et environnemental dès 2015 (Pérès, 2015). Pour les diffuseurs de données, l'éducation est une frontière indépassable.

Les stratégies d'éducation

Même si l'idée que le « grand public » puisse aller puiser dans les banques de données reste utopique, l'importance de diffuser une *data literacy*, une culture qui permette à chacun de lire, créer et communiquer des données, tout au moins d'en comprendre les enjeux, semble essentielle. L'un des objectifs du projet de recherche était d'étudier, en regard des pratiques déclarées des diffuseurs de données, les pratiques d'enseignants et de formateurs, de l'école primaire à l'université, recourant aux données ouvertes dans un objectif pédagogique avec leurs élèves et leurs étudiants. On constate que peu d'enseignants se sont encore saisis de ces ressources. Ceux qui l'ont fait ont mobilisé beaucoup d'énergie et ont immédiatement trouvé des réseaux d'échange (Travaux académiques mutualisés dédiés aux données ouvertes dans le domaine de l'éducation aux médias en 2017-2017 et des sciences économiques en 2017-2018, notamment). Bien qu'intéressés par le sujet, ils se confrontent à sa complexité. Certains éléments d'enquête, issus des entretiens individuels auprès d'enseignants, démontrent l'existence de réelles difficultés à identifier et définir les données ouvertes ainsi qu'à envisager leurs usages en situation d'enseignement-apprentissage.

Les données ouvertes constituent cependant un enjeu éducatif central autour du *big data* en vue de la maîtrise par les futurs citoyens des données qui les concernent personnellement, dans leurs activités ou dans leurs prises de décision, y compris sur le plan politique. Plusieurs compétences sont visées, permettant d'acquérir une culture analytique et critique des éléments de traitement, de communication et de réutilisation des données. Cette culture intègre des compétences complexes comme la capacité à trouver, nettoyer et exploiter des jeux de données, à produire des statistiques, à traduire des questions liées aux connaissances scolaires dans une datavisualisation (Capelle, Lehmans, Liquète, 2017), jusqu'à la compréhension des « algorithmes qui transforment et traduisent les données en visualisation » (Desfriches Doria, 2015, p. 54). Pour les enseignants, faire avec cette complexité des techniques, mais aussi des enjeux, demande beaucoup de temps, souvent trop par rapport à celui dont ils disposent pour former des « futurs citoyens » dans le cadre de leurs programmes. Les stratégies pédagogiques les plus élaborées sont basées sur la création de données nouvelles, dans une perspective de participation (Capelle, Lehmans, Liquète, 2017). Elles consistent à faire produire par les élèves eux-mêmes des données, en vue de la construction d'information puis de connaissance, en cartographiant par exemple des éléments du patrimoine, ou des données utiles aux déplacements des personnes en situation de handicap. La culture des données peut ainsi constituer une partie des « entourages qui procurent aux individus les conditions de leur liberté » (Zask, 2011, 58) dans un espace public où la participation relève d'une expérience concrète. Participer, comme le souligne Joëlle Zask avec John Dewey (2003), c'est prendre part à une enquête qui contribue à fabriquer le commun.

Conclusion

Les données sont un élément de base du dialogue ouvert qui se noue entre les collectivités publiques, les entreprises et les citoyens. Ces derniers sont des acteurs essentiels dans le processus d'ouverture des données, en tant qu'utilisateurs mais aussi acteurs potentiels d'une collaboration pour identifier les besoins ou interroger les processus, les espaces et les méthodes pour des usages sociaux et économiques diversifiés, dans un processus de co-construction ou de débats. Mais pour qu'ils puissent réaliser ce potentiel, un effort d'éducation et de formation reste à faire. Quelques enseignants en sont conscients, tout en soulignant la complexité d'une question qui tisse des enjeux politiques et techniques. Nous faisons le postulat que la valorisation des données ouvertes peut constituer un élément de "communs de la connaissance" à construire (Latour, 2011). L'analyse des processus d'ouverture montre que la mise à disposition de données relève de stratégies variées et parfois contradictoires. Dans les représentations des élus et des cadres, les enjeux sont essentiellement communicationnels, lorsqu'il s'agit de ne pas rester en dehors de la force d'attraction

du *big data*. Ils peuvent également être organisationnels, lorsqu'il s'agit de rénover le fonctionnement des services, de décloisonner les circuits fermés de l'information et de réviser leur articulation au regard du système d'information. Ils peuvent enfin relever d'un véritable projet politique de développement d'espaces publics de la donnée, de l'information et de la connaissance. L'espace public des données ouvertes est un espace à conquérir et à occuper pour devenir commun.

Références bibliographiques

Capelle, Camille, Lehmans, Anne, Liquete, Vincent (2017). « De la visibilité à la médiation : l'open data et ses usages en éducation ». Colloque international *Big data et visibilité en ligne, un enjeu pluridisciplinaire de l'économie numérique*, Novembre, Fort de France. URL : halshs01633284.

Chartron, Ghislaine, Broudoux, Évelyne (2015), « Enjeux géopolitiques des données, asymétries déterminantes », *in Big Data - Open Data : Quelles valeurs ? Quels enjeux ?* Actes du colloque « Document numérique et société », Rabat, De Boeck Supérieur, p. 65-83.

Damome, Étienne (2018), « Opportunités et difficultés du développement des archives ouvertes pour la communication publique : la situation en Afrique subsaharienne », *Revue française des sciences de l'information et de la communication* [En ligne], http://journals.openedition.org/rfsic/3491.

Desfriches Doria, Orélie (2015). « Quels dispositifs numériques pour appréhender la datavisualisation ? », *I2D – Information, données & documents*, vol. 52, n° 2, p. 54-56.

Dewey, John (2003), *Le public et ses problèmes*, trad. et introd. J. Zask, Pau : Farrago / Léo Scheer.

El Hachani, Mabrouka (2015), « Open data, collectivités et usagers : une dynamique en question », *in* Paquienséguy, Françoise (dir.), *Open data. Accès, territoires, citoyenneté : des problématiques info-communicationnelles,* Paris : Éditions des archives contemporaines, p. 1-23.

FRANCE. Secrétariat général pour la modernisation de l'action publique (2016), *Rapport d'activité 2015*, http://www.ladocumentationfrancaise.fr/rapports-publics/164000176/index.shtml

Goeta Samuel (2016), « Instaurer des données, instaurer des publics : une enquête sociologique dans les coulisses de l'open data ». Thèse en sociologie, Télécom ParisTech, URL: ⟨tel-01458098⟩

Goncalves, Dann, Rufat, Samuel (2016), « Open data et droit de la donnée : les collectivités à l'épreuve des réglementations européennes », *Cybergeo : European Journal of Geography* [En ligne], Science et Toile, document 787, mis en ligne le 17 août 2016, URL : http://journals.openedition.org/cybergeo/27750.

Labelle, Sarah, Le Corf, Jean-Baptiste (2012), « Modalités de diffusion et processus documentaires, conditions du « détachement » des informations publiques. Analyse des discours législatifs et des portails *open data* territoriaux», *in* Bardou Boisnier Sylvie et Pailliart Isabelle (coord.), Dossier *Information publique : stratégies de production, dispositifs de diffusion et usages sociaux*, Les Enjeux de l'Information et de la Communication, n° 13-2, p. 59-71, URL : https://lesenjeux.univ-grenoble-alpes.fr/pageshtml/art2012.html#dossier.

Latour, Bruno (2011), « Il n'y a pas de monde commun : il faut le composer », *Multitudes*, Vol. 45, n° 2, p. 38-41.

Lehmans, Anne, Capelle, Camille (2016), « Gouvernance et horizon d'attente des données ouvertes pour l'éducation », *Actes du 12ème colloque EUTIC*, Zakynthos, Grèce, 2016, URL : 〈hal-01790492〉 .

Lehmans, Anne (2018), « L'horizon d'une culture de la donnée ouverte : de l'utopie aux pratiques de gouvernance des données », *Revue COSSI*, n°1, URL : http://www.revue-cossi.info/numeros/n-1-2018-big-data-thick-data/708-1-2018-revue-lehmans#citer

Liquète, Vincent, Gilliard, Armelle (2017), « Épistémologie de la donnée au risque de la connaissance : approches croisées. », Communication au colloque COSSI *Méthodes et stratégies de gestion de l'information par les organisations : des "big data" aux "thick data"*, 11 et 12 mai 2017, Université McGill, Montréal, Canada.

Mabi, Clément (2015), « La plate-forme « data.gouv.fr » ou l'open data à la française », *Informations sociales*, Vol. 191, n° 5, p. 52-59.

Maday, Charlotte (2015), « L'apport de la gestion des documents d'activité (records management) à l'ouverture des données. Réflexions basées sur les pratiques en France », *Les cahiers du numérique*, Vol. 11, n° 2, p. 149-166.

Maurel Dominique, Chebbi Aïda (2012), « La perception de la confiance informationnelle. Impacts sur les comportements informationnels et les pratiques documentaires en contexte organisationnel », *Communication & Organisation*, vol. 2, n° 42, p. 73-90.

Meszaros, Branislav, Samath, Sitthida, Guérin-Hamdi, Sonia, Faure, Céline (2015), *Livre blanc sur les données ouvertes*, URL : <halshs-01162692>.

Pasquier, Martial, Villeneuve, Jean-Patrick (2007), « Les entraves à la transparence documentaire. Établissement d'une typologie et analyse des comportements organisationnels conduisant à empêcher ou à restreindre l'accès à l'information », *Revue Internationale des Sciences Administratives*, vol. 73, n° 1, p. 163-179.

Pérès, Éric (2015), « Les données numériques : un enjeu d'éducation et de citoyenneté », Avis du Conseil économique, social et environnemental, URL : http://www.lecese.fr/sites/default/files/pdf/Avis/2015/2015_01_donnees_numeriques.pdf

Peugeot, Valérie (2016), « Facilitatrice, protectrice, instituante, contributrice : la loi et les communs », Contribution au *colloque de Cerisy - Vers une république des biens communs ?*, Septembre 2016, URL : https://vecam.org/Facilitatrice-protectrice-instituante-contributrice-la-loi-et-les

Schafer, Valérie, Le Crosnier, Hervé (2011), *La Neutralité de l'Internet : une question de communication*, Paris : CNRS éditions.

Zask, Joëlle (2011), *Participer. Essai sur les formes démocratiques de la participation*, Lormont : Le Bord de l'eau.

Web sémantique : les politiques du sens et la rhétorique des données

Semantic Web: the politics of meaning and the rhetoric of data

Web Semántica: la política del significado y la retórica de los datos

Article inédit, mis en ligne le 15 novembre 2018.

Guillaume Sire

Maître de conférences en sciences de l'information et de la communication à l'Université Toulouse 1 Capitole et membre de l'Institut du Droit de l'Espace, des Territoires, de la Culture et de la Communication, et co-responsable de l'Unité Régionale de Formation à l'Information Scientifique et Technique d'Occitanie. Ses travaux portent sur la gouvernance du Web, les moteurs de recherche, le code informatique et la socio-économie des industries culturelles. / guillaume.sire@ut-capitole.fr

Plan de l'article

Résumé

Nous expliquons comment fonctionne le Web sémantique, et en particulier comment les producteurs de contenus peuvent « discrétiser » les informations de manière à ce que des logiciels comme les moteurs de recherche puissent ensuite les corréler à l'échelle du Web en générant des ontologies. Nous comparons les trois syntaxes auxquels ils peuvent avoir recours : le *Resource Description Framework*, les microdonnées et les microformats, ainsi que les arènes et les processus de normalisation propres à chacune de ces syntaxes. En nous référant pour notre analyse à la théorie opérationnelle de l'écriture numérique développée à l'Université Technologique de Compiègne, nous montrons ainsi comment sont confrontées différentes « politiques du sens » sur le Web portées par des acteurs dont les objectifs et les intérêts diffèrent. Nous décryptons la teneur non pas simplement du mécanisme de description et de traitement de l'information mais aussi des allers-retours entre l'individu qui décrit et édite des informations et la machine qui traite et éditorialise des données.

Mots clés

Gouvernance d'Internet, Web sémantique, ontologies, HTML, RDF, microdonnées.

Abstract

We explain how the semantic Web works, and in particular how content providers can "discretize" information in order to correlate them within ontologies. We compare three syntaxes they can use: *Resource Description Framework*, microdata and microformats, and we compare the normalization processes of each of these syntaxes. We show how different "politics of meaning" are confronted on the Web and supported by companies whose objectives and interests are far to be the same. We analyze not just the mechanism of description operated by hypertext markups but also the mechanism of interaction between the people who describe and publish content on the Web and the machines that process and editorialize data.

Keywords

Internet Governance, Semantic Web, Ontologies, HTML, RDF, microdatas.

Resumen

Explicamos cómo funciona la Web semántica y, en particular, cómo los productores de contenido pueden "discretizar" la información para que software como los motores de búsqueda puedan correlacionarlos en la Web generando ontologías. Comparamos las tres sintaxis que pueden usar: el Marco de descripción de recursos, los microdatos y los microformatos, y las arenas y procesos de normalización específicos de cada una de estas sintaxis. De esta manera, mostramos cómo diferentes "políticas de significado" en la Web son confrontadas por actores cuyos objetivos e intereses difieren. Y desciframos el contenido no solo del mecanismo de descripción y procesamiento de la información, sino también del recorrido de ida y vuelta entre el hombre que describe y publica la información y la máquina que procesa y editorializa los datos.

Palabras clave

Gobernanza de Internet, Web Semántica, Ontologías, HTML, RDF, microdatos.

Introduction

Le terme « *Big Data* » désigne un phénomène double du point de vue de l'information. L'épithète Big renvoie à la surcharge informationnelle, le terme « Data » aux données, c'est-à-dire à un flux d'informations non pas linéaire mais « discrétisé ». Cette discrétisation a lieu lorsque « *le contenu est inscrit en un langage constitué d'unités discrètes indépendantes les unes des autres* » (Bachimont 1999, p. 200). Discrétiser une information autrement dit revient à la « *transformer en une suite d'unités élémentaires (...) manipulables par une machine* » (Goyet, 2017, p. 68). Les « *Big Data* » supposent de traiter une quantité indénombrable d'unités qualifiables. Comment et à quelles conditions l'information ainsi multipliée et discrétisée par une succession de dispositifs sociotechniques peut-elle avoir du « sens » ? Et que devient dans ce cas la tension inhérente à toute technique, « *entre l'ouverture au sens et la fermeture sur une utilité fixée a priori* » (Bachimont, 2010, p. 24) ? Voilà des questions qui relèvent du champ des sciences de l'information et de la communication et dont les enjeux ne sont pas seulement techniques, mais aussi, et tout autant, culturels, politiques, économiques et juridiques.

Le Web a été le théâtre de la multiplication des informations, d'une part, et, d'autre part, de leur discrétisation. Tandis que la multiplication avait lieu parce que l'action de publier était désormais possible pour n'importe qui dès lors qu'on possédait de quoi se connecter, la discrétisation eut lieu étant donné le besoin de normaliser les pratiques de production, de sorte que les informations une fois publiées pourraient effectivement être traitées par les navigateurs chargés de les afficher, et par les moteurs chargés de les indexer. Une des étapes fondamentales de cette discrétisation consiste à diviser le contenant (la page) en fragments (titre, sous-titre, corps du texte, pied de page) identifiés par des balises de code HTML. L'autre étape fondamentale consiste à discrétiser le contenu (le texte) en trouvant le moyen de faire voyager non pas seulement des signaux mais aussi des signes. Par exemple, pour un nombre, on prévient les navigateurs et les moteurs de recherche qu'il s'agit d'une taille, d'un poids, d'une date, etc. En se référant à la théorie opérationnelle de l'écriture numérique de Crozat *et al.* (2011), on peut dire qu'au lieu de se contenter de prendre en charge le « niveau techno-applicatif », en transformant les signaux binaires (une suite de 0 et 1) en nombre (mettons « 48 »), la machine prend également en charge le « niveau sémio-rhétorique » en attachant à ce nombre un sens (taille, poids, etc.). A la suite de ce processus, ce qui est manipulable correspond à ce qui est signifiant, la syntaxe ayant permis d'opérer une « *coïncidence entre le format technique et la forme sémiotique* » (Bouchardon *et al.*, 2011, p. 14). C'est ce qu'on appelle le Web sémantique. En plus de donner une structure à l'information, en discrétisant son contenant, on lui donne un sens, en discrétisant son contenu, puis on donne une structure aux unités de sens : une ontologie (Gruber, 1993).

Il existe plusieurs procédés de discrétisation sémantique. Chacun est normalisé de façon plus ou moins souple et transparente. Il y a différents codes (HTML et Javascript) et au sein d'un même code il peut y avoir différentes syntaxes utilisées pour transformer le flux d'information en unités signifiantes (RDFa, microdonnées et microformats pour le HTML, JSON-LD pour le Javascript), et différents formats de description ontologique utilisés pour structurer les relations entre ces unités (RDFS, OWL). Les conditions de production du sens — c'est-à-dire les conditions devant être remplies pour qu'il puisse y avoir « déprise et reprise » des unités discrètes (Bachimont, *op. cit.*, p. 29) — dépendent des spécificités de chacun de ces codes et de ces syntaxes, de leurs auteurs et de leurs modes de gouvernance. Les sciences de l'information et de la communication ont donc intérêt à investiguer la question de la concurrence et de la complémentarité entre ces codes, ces syntaxes et ces ontologies, et à analyser de quelles valeurs chacun d'entre eux est le véhicule et le garant. Cela car comme l'avait souligné Lawrence Lessig dans son célèbre article — *Code is Law* (1999) — les choix concernant le code sont des choix de valeurs. En comprenant les tenants des choix possibles en matière de Web sémantique, on comprendra non pas seulement les conditions de production du sens, du point de vue technique, mais vraiment ce que nous pourrions appeler « les politiques du sens ». On comprendra quels sont les intérêts qui prévalent, qui sont les acteurs qui dominent, et quels sont les effets concrets de cette prévalence et de cette domination en termes d'information et de communication. En outre, on comprendra où est *située* la sémantisation, en sondant non pas simplement le mécanisme de description et de traitement de l'information mais aussi la teneur des allers-retours entre l'homme qui décrit le contenu et la machine qui le traite.

Cadrage théorique et cadre méthodologique

Les syntaxes que nous étudierons ici sont des normes, en cela qu'elles se présentent toutes sous la forme de documents « *déterminant des spécifications techniques de biens, de services ou de processus qui ont vocation à être accessibles au public, [qui] résultent d'un choix collectif entre les parties intéressées à sa création, et [qui] servent de base pour la solution de problèmes répétitifs* » (Lelong et Mallard, 2000, p. 11). C'est pourquoi nous nous sommes placés dans la même perspective que ceux qui parmi nos collègues intéressés par la gouvernance de l'Internet étudient en particulier la normalisation des protocoles (De Nardis, 2009 ; Ermoshina et Musiani, 2018) en partant du principe qu'étudier comment une norme est construite et circule revient à étudier des rapports de pouvoir (Bowker et Star, 1999 ; Bush, 2011).

D'autre part, dans notre étude, l'information et la communication ne sont pas des variables explicatives, l'*explanans*, mais des variables expliquées : l'e*xplanandum.* Nous voulons comprendre pourquoi les syntaxes sont utilisées comme ça et pas autrement, et pourquoi la discrétisation de

l'information qu'elles opèrent a lieu de cette façon et pas d'une autre. C'est pourquoi nous nous sommes appuyés sur la théorie du support comme elle a été développée en particulier à l'Université Technologique de Compiègne par les chercheurs Bruno Bachimont, Serge Bouchardon, Isabelle Cailleau et Stéphane Crozat. En développant une théorie opérationnelle de l'écriture numérique, ceux-là se sont intéressés à la manière dont le numérique peut sortir de « l'autarcie » pour être « *négocié dans le monde de la matière et du sens* » (Bouchardon *et al.*, 2011, p. 13). Notre étude vise justement à comprendre de quoi est faite cette négociation dans le cas du web sémantique : comment, concrètement, les possibilités d'écriture sont négociées, en gardant à l'esprit que la compréhension de cette négociation éclairera les conditions *effectives* de production (information) et de mise en circulation (communication) du sens.

Pour comparer les organismes de normalisation, leurs prises d'intérêts et leurs procédures, nous avons commencé par identifier ceux qui tenaient les rênes des trois principaux organismes concernés par la standardisation des protocoles du Web sémantique (*World Wide Web Consortium*, *CommerceNet* et *Web Hypertext Application Technology Working Group*). Puis nous avons analysé comment au sein de ces organismes les décisions étaient prises. Enfin, nous avons comparé les modalités d'utilisation des différentes syntaxes du Web sémantique. Pour cela, nous avons nous-même produit du code. Cela nous a obligé à apprendre à maîtriser chacune des syntaxes de façon à être capable de comprendre ce que l'utilisation de l'une d'entre elles impliquait par rapport à l'utilisation d'une autre. Plus précisément, nous avons exprimé une même recette de cuisine dans les cinq principales syntaxes du Web sémantique (RDF, RDFa, Schema.org, hRecipe, JSON-LD). Cela nous a permis de représenter un objet dans différents langages de manière à pouvoir analyser et comparer ce que chacune de ces représentations faisait à/de l'objet (la recette de cuisine). Nous avons ainsi pu identifier non pas seulement les tenants de la concurrence entre syntaxes mais aussi les aboutissants sémiotiques de cette concurrence, *en les expérimentant*.

Nous nous sommes concentrés sur les syntaxes qui opèrent la discrétisation en structurant le contenu du document, et permettent de constituer des bases de données qui ensuite sont structurées grâce à des ontologies. Nous n'avons donc pas analysé les formats de description ontologique (dont les deux principales sont RDFS et sa spécification OWL), lesquels agissent une fois seulement que l'information a été discrétisée, et qui sont donc plus éloignées des créateurs de pages Web qui n'y ont jamais recours eux-mêmes. A l'inverse, tous les individus qui publient des informations sur le Web et décident d'opérer un balisage sémantique doivent obligatoirement choisir entre les différentes options que nous décrirons ci-après.

W3C et RDF/RDFa

Le *World Wide Web Consortium* ou « W3C » a été créé par le fondateur du Web Tim Berners-Lee. C'est une organisation à but non lucratif ouverte à toutes les personnes morales qui le souhaitent. Il leur suffit pour devenir membre de payer une somme pouvant aller d'un montant très modique pour une association, jusqu'à 59 000 euros par an pour une grande entreprise. Ils étaient au total 463 membres en septembre 2017. C'est au W3C que sont édités les trois standards mis au point par Tim Berners-Lee au début des années 1990 : l'*Uniform Resource Locator* (URL) pour localiser les données, l'*HyperText Transfer Protocol* (HTTP) pour les transférer et l'*HyperText Markup Language* (HTML) pour les décrire. Chaque proposition de modification ou de création d'un standard fait l'objet d'une discussion entre les membres du W3C ayant émis le souhait de participer au groupe de travail concerné. Elle donne ensuite lieu à un document censé franchir quatre étapes à l'issue desquelles, s'il les a toutes franchies, le document en question obtient le statut de norme officielle. C'est durant ces étapes que se joue ce qu'en sociologie des techniques on nomme la « stabilisation » (Bijker, 1995) et la « structuration » des relations technologiques (Benezech, 1996). Le franchissement de chacune d'elles se fait par consensus, sachant que c'est Tim Berners-Lee, ou une personne qu'il aura expressément mandatée pour ce faire, qui décrète qu'il y a consensus sur telle ou telle question et qu'il est temps de passer à l'étape suivante, ou qui décrète au contraire que le consensus n'a pas été trouvé et qu'il convient de revenir à une étape précédente, ou tout simplement d'abandonner le projet. Le processus du W3C confère un pouvoir considérable à Tim Berners-Lee, qui peut prendre en dernière instance des décisions contraires à l'avis exprimé par la

majorité des membres d'un groupe de travail, ou bien trancher alors qu'il reste des objections formelles exprimées par tel ou tel membre et n'ayant pas été résolues (Halpin, 2017 ; Sire, 2017). Pour cette raison, certains observateurs n'hésitent pas à qualifier le fondateur du Web de « dictateur à perpétuité » (Malcom, 2008). Il n'empêche, le W3C est une arène dont la vocation est d'être ouverte, transparente, de garantir l'efficacité technique des normes, d'assurer l'interopérabilité des infrastructures (Virili, 2003), et d'empêcher, grâce à une politique de brevets originale, que des entreprises participent à la discussion dans le seul but de se rendre indispensables par la suite en tirant parti des brevets qu'elles détiennent (Russel, 2003 ; Halpin, 2017).

Au sein du W3C, le groupe de travail intéressé par les langages sémantiques s'appelle le *Semantic Web Interest Group* (SWIG). C'est en son sein qu'a été développée la syntaxe RDF, basé sur le code XML, et ayant atteint le statut de standard officiel en 2004 en version 1.1. Si on observe cette syntaxe au prisme de la théorie opérationnelle de l'écriture, on y retrouve le tropisme de « manipulabilité » : il est possible d'appliquer aux données discrétisées des règles formelles de traitement algorithmique. On retrouve également le principe de « fragmentation-agrégation » : le contenu est fragmenté en unités potentiellement signifiantes qu'on pourra éventuellement agréger différemment (Crozat *et al.*, 2011, p. 21-22).

Il est important de comprendre pour la suite que le RDF est construit de telle sorte que soient renseignés des triplets sous la forme [sujet {prédicat} objet]. Par exemple pour la phrase : « *Octave a cuisiné un gâteau au chocolat pour Romane* », nous aurons les trois prédicats suivants :

- [Octave {est l'auteur du} gâteau au chocolat]
- [Romane {est le propriétaire du} gâteau au chocolat]
- [Le gâteau {est au} chocolat]

Nous pouvons créer d'autres relations grâce à des ontologies : si Octave est l'auteur du gâteau au chocolat et Romane son propriétaire, alors : [Octave {est un fournisseur de} Romane]. Il s'agira d'activer cette fois ce que la théorie opérationnelle de l'écriture nomme « la fonction *génération* » : on pourra créer « *des contenus automatiquement à partir des contenus existants* » (Crozat *et al.*, 2011, p. 22). Les sujets et les objets peuvent être des personnes, des choses, des concepts, des dates, des lieux, ou bien des adresses URL. Si on parvient à donner à chaque « fragment » (personnes, choses, concepts, dates, lieux...) un seul *Uniform Resource Identifier* (URI), dans une base de données où sont renseignées les triplets (qu'on appelle « dépôt » ou « banque »), il devient possible de créer des ontologies à l'échelle du Web. Après avoir transformé les informations continues en données discrètes (*manipulabilité*), on met de l'ordre non plus entre les contenants (i.e. en hiérarchisant les sites et les pages) mais entre les contenus (*fragmentation-agrégation*), c'est-à-dire entre les données elles-mêmes, archivées dans différentes ontologies interconnectées, à partir desquelles on pourra générer de nouveaux contenus (*génération*).

Pour employer le RDF, il faut utiliser un schéma de métadonnées permettant aux développeurs de décrire les éléments de leurs pages Web de façon à être compris par les logiciels type moteurs de recherche, lesquels pourront ensuite générer eux-mêmes des ontologies, ou bien se référer à une ontologie existante comme DBpédia (collection de triplets constitués à partir des informations contenues dans Wikipedia). C'est un cas typique du niveau « techno-applicatif » : il s'agit de proposer « un formatage qui prescrit *a priori* des signes et leur manipulation formelle » (Crozat *et al.*, *op. cit.*, p. 17). Il existe plusieurs schémas ou « espaces de noms » utilisés en RDF et correspondant à chaque fois à des ontologies spécifiques, par exemple « Dublin Core » et « Friend Of A Friend » qui peuvent être utilisés en plus de l'espace de nom RDF (RDFS). Certains de ces schémas, comme le Dublin Core, existait avant le RDF. Ainsi, le RDF permet de faire le lien avec des techniques de discrétisation préexistantes. Parce que plusieurs schémas peuvent être utilisés au sein d'un même document, il est indispensable de renseigner les préfixes à chaque fois dans les balises XML tout en ayant annoncé en tête du document quels étaient les différents schémas utilisés et à quelle adresse le logiciel pourra en trouver une description (on se sert pour cela du préfixe « *xmlns* » pour « *xml name space* »). Ci-dessous, nous avons écrit une recette de gâteau au chocolat, puis nous avons ensuite discrétisée en code XHTML (XML + HTML), grâce à la syntaxe RDF :

Version brute :

Recette du gâteau au chocolat d'Octave
Dessert pour 4 personnes / 15 minutes de préparation / 1er février 2018
Ingrédients :
200 g de chocolat noir à 70% de cacao
125 g de beurre demi-sel
100 g de farine
1 sachet de levure chimique (10 g)
4 oeufs
200 g de sucre en poudre
1 pincée de sel
Préparation :
Coupez le chocolat et le beurre en morceaux. Faites fondre au bain-marie. Retirez du feu.
Ajoutez la farine et la levure tamisées dans le chocolat fondu.
...

Version RDF :

```
1. <html xmlns=http://www.w3.org/1999/xhtml
2. xmlns:foaf="http://xmlns.com/foaf/0.1/"
3. xmlns:re="http://octave.com/recettes/gateaux">
4. <head>
5. <title>Recette du gâteau au chocolat d'Octave</title>
6. </head>
7. <body>
8. <h1> Recette du gâteau au chocolat d'Octave </h1>
9. <re:recipe>
10. <re:recipe_head>
11. <foaf:name>Octave</foaf:name>
12. <re:meal_type>Dessert</re:meal_type>
13. </re:recipe_head>
14. <re:recipe_body>
15. <re:ingredients>
16. <re:ingredient> 200 g de chocolat noir à 70% de cacao </re:ingredient>
17. <re:ingredient> 125 g de beurre demi-sel</re:ingredient>
18. <re:ingredient> 100 g de farine</re:ingredient>
19. <re:ingredient> 1 sachet de levure chimique (10 g)</re:ingredient>
20. <re:ingredient> 4 oeufs </re:ingredient>
21. <re:ingredient> 200 g de sucre en poudre</re:ingredient>
22. <re:ingredient> 1 pincée de sel</re:ingredient>
23. <re:directions>
24. <re:direction>Coupez le chocolat et le beurre en morceaux. Faites-les fondre en bain-marie.
    Retirez du feu. </re:direction>
25. <re:direction>Ajoutez la farine et la levure tamisées dans le chocolat fondu.</re:direction>
26. ...
27. </re:directions>
```

```
28. </re:recipe_body>
29. <re:recipe_footer>
30. <re:serving>4</re:serving>
31. <re:preparation_time>15 minutes</re:preparation_time>
32. </re:recipe_footer>
33. <re:document_info>
34. <re:document_author>Octave</re:document_author>
35. <re:date_updated>01/08/2018</re:date_updated>
36. </re:document_info>
37. </re:recipe>
38. </body>
39. </html>
```

Nous avons fait référence à deux espaces de noms (lignes 2 et 3). Le premier : *Friend of a Friend* (http://xmlns.com/foaf/0.1/) auquel nous avons donné le préfixe « foaf ». Et un deuxième que nous avons créé à titre d'exemple (http://octave.com/recettes/gateaux) auquel nous avons donné le préfixe « re » comme « recette ». Les balises spécifiques à chaque espace de nom sont assorties du préfixe correspondant, de telle sorte que les navigateurs et les moteurs de recherche puissent savoir que c'est bien à cet espace de nom, et aux ontologies qui lui sont propres, que la balise renvoie. Le principe de *fragmentation*, attaché au tropisme de *manipulabilité*, donne ainsi lieu à des possibilités de *génération* d'autant plus nombreuses que plusieurs espaces de noms pourront être invoqués.

Dans le HTML5, on utilise le *Resource Description Framework in Attributes* (RDFa) normalisé par le W3C en 2008 pour la version 1.0 et en 2013 pour la version 1.1. Il fonctionne comme le RDF mais les préfixes cette fois s'inscrivent dans les attributs des balises HTML au lieu de s'inscrire dans le nom des balises. Cela simplifie considérablement l'écriture, d'une part, et, d'autre part, cela permet de mieux normaliser l'usage des balises et de faire varier seulement les attributs à l'intérieur des balises. On augmente ici la *manipulabilité*, de sorte que les développeurs pourront opérer une *fragmentation* plus fine, ce qui augmentera l'efficacité des dispositifs visant à *générer* de nouveaux types de contenus à partir d'ontologies. Le RDFa utilise les attributs HTML existants (notamment class, id, rel, rev et href) et en institue de nouveaux (about, property, content, datatype, resource, typeof, prefix, vocab), mais les balises restent les mêmes quel que soit l'espace de nom mobilisé. La fragmentation, autrement dit, se joue à un niveau inférieur de la syntaxe : ce ne sont plus les éléments syntaxiques qui varient, mais les attributs d'éléments qui sont les mêmes dans tous les cas, ce qui permet d'optimiser le paramétrage des logiciels chargés de comprendre et de traduire les informations ainsi transformées en données (navigateurs, moteurs de recherche). Pour notre recette, cela donne :

```
1. <p xmlns:re=http://octave.com/recettes/gateaux
   xmlns:foaf="http://xmlns.com/foaf/0.1/">
2. <head>
3. <title>Recette du gâteau au chocolat d'Octave</title>
4. </head>
5. <body>
6. <h1> Recette du gâteau au chocolat d'Octave </h1>
7. <span property="re:recipe_author" about="foaf:name">Octave</span>
8. <span property="re:meal_type">Dessert</span>
9. <span property="re:ingredient"> 200 g de chocolat noir à 70% de cacao </span>
10. <span property="re:ingredient"> 125 g de beurre demi-sel</span>
11. ...
```

Les balises sont des balises HTML classiques mais les « xmlns » renseignés à la première ligne donnent lieu à des préfixes à l'intérieur des attributs, c'est-à-dire à l'intérieur des balises, de sorte qu'une même balise puisse contenir plusieurs attributs associés chacun à un schéma de métadonnées, et à une ontologie, spécifiques (cf. ligne 7). La discrétisation est plurielle : une même unité de sens renvoie à plusieurs modes de représentation.

Le RDF et le RDFa ne sont pas les seules syntaxes permettant d'opérer la discrétisation. Il y en a deux autres, les microformats et les microdonnées, qui chacune ont à la fois leurs propres modes de gouvernance, et leurs propres principes de fragmentation-agrégation, ayant eux-mêmes des effets sur les possibilités de génération de contenus (i.e. les moteurs de recherche n'effectueront pas les recoupements ontologiques *dans les mêmes conditions*), et, donc, sur le niveau « sémio-rhétorique ».

CommerceNet et Microformats

Les microformats (auxquels renvoient les sigles µF ou uF) ont été développés sans politique de normalisation arrêtée, de façon participative. Cela les différencie des syntaxes RDF/RDFa normalisées selon la procédure du W3C. Le fait que les microformats n'aient jamais été « stabilisés » les apparente à ce que Ksenia Ermoshina et Francesca Musiani nomment un « quasi-standard » (Ermoshnia et Musiani, 2018). N'importe qui peut en théorie proposer un nouveau type de microformat en créant son propre espace de noms et sa propre syntaxe. Il faut pour cela suivre une procédure qui consiste essentiellement à ouvrir une discussion sur le wiki de CommerceNet. Sur ce même site, il est dit qu'avant de proposer un nouveau type de microformat il faut implémenter sur son site les microformats qui existent déjà, comme un gage de bonne foi, et publier un « témoignage ». Mais toutes les propositions ne seront pas implémentées, et seules celles qui créeront autour d'elles un consensus sans que personne n'ait le pouvoir de *décréter* que c'est le cas, atteindront le statut de « quasi-standard ».

L'organisation à but non lucratif CommerceNet, dont la vocation est de promouvoir le commerce électronique, a aidé à mettre en place les microformats et à fédérer une communauté autour du wiki « Microformats.org ». La documentation officielle prévient que les microformats sont « conçus d'abord pour les humains, ensuite pour les machines » (Khare et Çelik, 2006). L'accent est mis tout au long de la procédure informelle pour que les microformats soient lisibles pour des êtres humains, c'est-à-dire suffisamment proches du langage naturel pour être compréhensibles y compris en affichant le code source d'une page pour la lire à l'œil nu. Contrairement au RDF/RDFa, et en théorie (nous verrons qu'en pratique c'est différent), les microformats se situent au niveau sémio-rhétorique plutôt qu'au niveau techno-applicatif : ils ont vocation selon leurs concepteurs à créer du sens en interagissant avec des humains, et, dans un second temps seulement, à donner lieu à des traitements logiciels, en interagissant avec des machines.

Les microformats reposent sur trois attributs (class ; rel ; rev). Ces attributs peuvent être insérés dans n'importe quelle balise HTML et, dans le cas où il n'y en a pas déjà une à l'endroit où il faut insérer l'attribut, ils peuvent être ajoutés grâces aux balises ou . Tous les microformats existants sont répertoriés sur le wiki (http://microformats.org/wiki/Main_Page). Chacun est associé à un type d'information (description d'une personne, localisation d'un lieu, CV, petites annonces) et assorti d'une page Web présentant aux développeurs comment l'utiliser. Il suffit d'annoncer à quel « microformat » on se réfère dans une balise HTML située en amont du document et de suivre les instructions du wiki pour renseigner les balises situées en aval. Par exemple, pour le microformat « hrecipe », nous avons :

```
1. <div class="hrecipe">
2. <h1 class="fn"> Recette du gâteau au chocolat d'Octave</h1>
3. Contributed by <span class="author">Octave</span>
4. <h2>Ingredients</h2>
5. <ul>
6. <li class="ingredient">
7. <span class="value">"> 200 g de chocolat noir à 70% de cacao </span>
8. ...
9. <h2>Instructions</h2>
10. <ul class="instructions">
11. <li> Coupez le chocolat et le beurre en morceaux. Faites-les fondre en bain-marie. Retirez
    du feu. </li>
12. ...
```

Le nom du microformat (hrecipe, ligne 1) joue le rôle joué par l'espace de nom dans RDF/RDFa. Mais contrairement aux espaces de noms, aucun lien n'est effectué vers le lieu où la syntaxe est décrite. La fragmentation n'induit pas la même manipulabilité, et ne suppose pas de générer du contenu aussi aisément qu'avec RDF/RDFa. Si le logiciel ne connaît pas déjà la syntaxe dont il est question (i.e. si son concepteur ne l'a pas paramétré en fonction d'une syntaxe donnée *a priori*), il n'aura aucun moyen de « comprendre » de quoi il est question (niveau sémio-rhétorique) ou de « générer » de nouveaux contenus (niveau techno-applicatif). Ainsi les microformats ne permettent-ils pas de faire le lien avec d'autres vocabulaires (dublin core, friend of a friend, rdf), pas plus qu'ils ne permettent de faire le lien avec les ontologies constituées sur la base de ces vocabulaires. Il y a discrétisation, certes, mais il faut que le logiciel effectue certaines opérations sur ces données pour pouvoir faire le lien avec d'autres données situées ailleurs sur le Web.

Les microformats permettent aussi d'éditorialiser l'affichage dans les listes de moteurs de recherche. C'est le cas notamment depuis que Google a proposé en 2009 d'utiliser les microformats pour structurer l'apparence des recettes de cuisine dans les listes de résultats. A la suite de cette annonce et de sa mise en œuvre, Google a créé « *Recipe View* » en 2011, un module spécifique permettant notamment aux usagers du moteur de spécifier les ingrédients dont ils disposent pour savoir quel plat il leur est possible de cuisiner. De ce point de vue, les microformats auraient moins vocation à opérer une sémantisation qu'à éditorialiser l'affichage sur les moteurs de recherche. L'intention de ceux qui s'en saisissent serait alors moins située au niveau sémio-rhétorique qu'au niveau techno-applicatif, et ce alors même que, nous l'avons vu, l'intention de ceux qui ont conçu les microformats, elle, est au contraire davantage située au niveau sémio-rhétorique (« il faut pour voir être compris par des êtres humains...), qu'au niveau techno-applicatif (... et éventuellement pouvoir être opéré par des machines calculatoires »). On peut émettre l'hypothèse forte selon laquelle cette ambivalence est peut-être due au fait que les microformats sont implémentés par des entreprises de e-commerce essentiellement, qui ont besoin d'outils pour mettre en valeur leurs produits dans les listes de résultats des moteurs de recherche, mais qui n'ont pas forcément envie que des ontologies soient

constituées qui permettraient de comparer les prix et les caractéristiques des produits à vendre à grande échelle.

Whatwg et Microdonnées

Le Whatwg a été créé à la suite d'un désaccord au sein du W3C. Une scission a eu lieu en 2005, lorsque Ian Hickson a proposé de travailler sur une version plus interactive du langage HTML, le HTML5, mais que Tim Berners-Lee a refusé d'ouvrir un groupe de travail, cela parce qu'il préférait se concentrer sur le XHTML et sur l'objectif de sémantiser le Web (Sire, 2017). Ian Hickson ne renonça pas à son projet. Il collabora avec des ingénieurs d'Apple, Mozilla et Opera, puis Google à partir de 2005 (qui embaucha Ian Hickson) pour créer le *Web Hypertext Application Technology Working Group* (Whatwg). Ce groupe de travail fut doté d'une procédure extrêmement souple comparée à celle du W3C. L'objectif était de normaliser une version du code HTML en dehors du W3C, ce qui était possible dès lors que les premiers soutiens du Whatwg étaient les propriétaires des navigateurs et des moteurs de recherche et pouvaient par conséquent paramétrer leurs logiciels de manière à exécuter les balises qu'ils auraient eux-mêmes mises au point, en plus de celles qui auraient été discutées et validées par le W3C (Sire, 2017).

Finalement, en 2007, le XHTML 2.0 était l'objet de plusieurs dysfonctionnements techniques, en conséquence de quoi Tim Berners-Lee proposa aux membres du Whatwg d'ouvrir un groupe de travail au sein du W3C concernant le HTML5, ce que Ian Hickson et ses collaborateurs acceptèrent sans pour autant mettre un terme à l'activité du Whatwg, disposant ainsi de deux structures où discuter du même protocole. Les navigateurs et les moteurs de toute manière reconnaîtraient les balises du code HTML5, et, ce, qu'elles émanent du W3C ou bien du Whatwg. Les deux arènes éditeraient le langage conjointement, chacune selon sa propre procédure, et les concepteurs de navigateurs participeraient aussi bien à l'une qu'à l'autre, discutant entre eux au Whatwg et avec d'autres au W3C mais pouvant décider, dans le cas où une recommandation du W3C ne leur conviendrait pas, de ne pas l'implémenter et de lui préférer des modalités issues du seul Whatwg dans le cas où un arbitrage devrait avoir lieu.

Concernant la procédure de normalisation, n'importe qui peut théoriquement participer aux discussions du Whatwg. Cependant le dernier mot appartient au « Steering Group » composé en 2018 d'un représentant pour chacune de ces quatre entreprises : Mozilla, Microsoft, Google et Apple. D'autre part, alors qu'elle n'en avait pas jusque-là, le Whatwg a mis en place en janvier 2018 une politique de brevet similaire à celle du W3C, basée sur la gratuité et sur l'engagement pour chaque contributeur à ne pas tirer parti financièrement d'un brevet nécessaire à l'implémentation du standard (Russel, 2003).

C'est dans le cadre du Whatwg qu'ont été développées les microdonnées. Il est important ici de bien considérer que ce sont les propriétaires de navigateurs et de moteurs de recherche, qui ont mis en place ce format. Il repose sur le principe du « *Living Standard* » : des modifications peuvent être faites rapidement, sans avoir à suivre les étapes et le protocole plus rigide des normes stabilisées par le W3C. Les microdonnées fonctionnent à partir d'attributs spécifiques : itemscope, itemtype, itemid, itemprop, itemref. L'attribut itemtype permet de renvoyer à un vocabulaire faisant office d'espace de nom, et présent sur le site « Schema.org ». Par exemple ci-dessous, il s'agit d'une recette de cuisine dont la première ligne nous apprend qu'elle utilise le vocabulaire décrit ici : « http://schema.org/Recipe ». En revanche, on ne peut pas renvoyer à d'autres vocabulaires qu'à ceux du Shema.org. Il y a fixation des balises et du vocabulaire, les variations ne concernant cette fois que les attributs, alors qu'elles concernaient la syntaxe et le vocabulaire dans le cas du RDF, et les attributs et le vocabulaire dans le cas du RDFa et des microformats. Autrement dit, le Schema.org, basé sur le tropisme de manipulabilité, limite le principe de fragmentation-agrégation pour mieux contrôler la fonction de traitement automatique et de génération des contenus.

```
1.  <div itemscope itemtype="http://schema.org/Recipe">
2.    <span itemprop="name">Recette du gâteau au chocolat d'Octave</span>
3.    By <span itemprop="author">Octave</span>,
4.    Prep Time: <meta itemprop="prepTime" content="PT15M">15 minutes
5.    Ingredients:
6.    - <span itemprop="recipeIngredient">125 g de beurre demi-sel</span>
7.    - <span itemprop="recipeIngredient">100 g de farine</span>
8.    - <span itemprop="recipeIngredient">1 sachet de levure chimique (10 g)</span>
9.    ...
10.   Instructions:
11.   <span itemprop="recipeInstructions">
12.   Couper le chocolat et le beurre en morceaux
13.   </span>
14. ...
15.   </div>
```

On voit ici que la discrétisation avec des balises qui dans presque tous les cas (sauf ligne 4) ont recours à la même balise () et au même attribut (itemprop), pour faire recours à un schéma exclusif, sur lequel aucune ontologie publique ne repose, et qu'il faudra donc « traduire » en RDF si l'on veut faire le lien avec les banques de triplets du type DBpédia (http://schema.org/docs/datamodel.html). Les microdonnées constituent donc un dispositif rendant plus simple la discrétisation au niveau techno-applicatif (peu de balises, peu d'attributs, un schéma très facile à comprendre et à utiliser), mais rendant plus compliquée la sémantisation au niveau sémio-rhétorique (on ne peut pas créer d'ontologie sur la seule base des unités de sens exprimées sous forme de microdonnées).

En 2011, après l'officialisation des microdonnées, Tantek Celik, l'un des créateurs des microformats, a accusé sur son compte Twitter le Whatwg de « cracher dans les yeux de toutes les personnes et les organisations ayant œuvré à la conception des vocabulaires ouverts vCard, iCalendar, etc. ». Le développeur (très influent) Mark Pilgrim a quant à lui prévenu que le Schema.org était l'illustration de l'échec du W3C avec son RDF/RDFa. Manu Sporny, qui dirigeait le groupe de travail du W3C autour de la spécification RDF, a pointé du doigt le fait que les microdonnées étaient le fait d'un petit groupe d'organisations alors que les RDFa et les microformats étaient le fruit du travail collectif de milliers de personnes, et que par conséquent les microdonnées ne pouvaient pas être considérées comme étant de véritables normes ouvertes. La dispute a été virulente, et les arguments avancés pour l'une ou l'autre des trois possibilités visaient tous à acquérir une légitimité plus forte que les autres, soit en prétendant que la syntaxe défendue était plus efficace techniquement et plus facile à généraliser, soit en attaquant le fonctionnement des arènes de normalisation concurrentes.

En plus d'être le seul vocabulaire à être utilisable avec la syntaxe des microdonnées, le « shema.org » peut être utilisé comme vocabulaire avec les syntaxes RDF et RDFa. Il permet ainsi de faire le lien avec les ontologies appuyées sur ces syntaxes. Finalement, du côté syntaxe (au niveau techno-applicatif) les microdonnées concurrencent les microformats, et du côté vocabulaire (au niveau sémio-rhétorique) le shema.org concurrence les vocabulaires utilisés en RDF/RDFa.

Le Whatwg a également proposé d'intégrer son vocabulaire à une syntaxe pour Javascript standardisée par le W3C : le « *JavaScript Object Notation for Linked Data* » (JSON-LD). Le principe cette fois n'est plus de discrétiser directement l'information présente dans le document mais *de la recopier en la discrétisant* dans un espace réservé aux logiciels, en la traduisant dans un code de programmation (le Javascript) moins proche de l'écriture que ne l'est le HTML (qui est un code de description). Tout en restant dans le niveau techno-applicatif, on remonte ainsi vers le niveau théorico-idéal, « celui du code binaire défini dans son double isolement interprétatif et matériel » (Crozat *et al.*, 2011, p. 16), et, donc, en s'éloignant du niveau sémio-rhétorique. Le JSON-LD se

différencie du RDF et des microformats, où l'être humain passait avant la machine, et où la syntaxe de la discrétisation était fondue dans la syntaxe de la langue naturelle, comme une espèce de ponctuation sémantique. Il aboutit à la création d'un espace « plus abstrait » que seule la machine a vocation à consulter, tandis que les internautes continueront à consulter sur leurs navigateurs l'information contenue dans un fichier débarrassé des indications sémantiques, et qui, si on le recopie et qu'on le transporte ailleurs, « voyagera » sans sémantisation.

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Recipe",
  "author": "Octave",
  "cookTime": "PT15M",
  "recipeIngredient": [
    "1 sachet de levure chimique (10 g)",
    ...
  "recipeInstructions": "Couper le chocolat et le beurre en morceaux",
...
</script>
```

L'usage du JSON-LD est recommandé par Google, car il est idéal pour le fonctionnement de son moteur de recherche. Il permet notamment à l'entreprise de créer sa propre ontologie, le *Knowledge Graph*, à partir des dépôts de triplets existants et des données que le moteur récolte lui-même, et de l'exploiter dans les résultats de son moteur de recherche. Mais cette ontologie, même si n'importe quel internaute peut l'interroger via le moteur, n'est pas accessible en tant que telle, ni réutilisable par une autre technologie. En devenant plus manipulables pour les machines, les données le sont moins, ou plus du tout, pour les êtres humains. Surtout, elles sont stockées dans un autre espace que celui où se trouvent les informations destinées aux humains. Le niveau techno-applicatif se dédouble : pour un même fond, on trouve d'un côté une forme adressée aux logiciels dont la vocation est d'agréger les contenus et d'en générer de nouveaux (les moteurs de recherche), et de l'autre une forme adressée aux logiciels dont la vocation est de structurer l'affichage du contenu sur l'écran de l'utilisateur (les navigateurs). Ainsi, Google, après avoir proposé, et imposé, un nouveau vocabulaire, le Schema.org, et une nouvelle syntaxe pour le code HTML, les microdonnées, a proposé de traduire dans un autre code cette syntaxe et ce vocabulaire, pour mieux fluidifier la relation entre les niveaux théorético-idéal et techno-applicatif, quitte à opérer une rupture avec le niveau sémio-théorique.

Conclusion : l'industrialisation du sens

Nous avons vu comment la discrétisation des informations présentes sur le Web, c'est-à-dire leur transformation en *Big Data*, pouvait donner lieu à trois manières très différentes de tisser des liens entre « les trois niveaux du numérique » : théorético-idéal, techno-applicatif, sémio-rhétorique (Crozat *et al,* 2011*)*. Il y a là trois « politiques du sens ». En comparant les procédures de normalisation, nous avons montré que dans le cas des microdonnées et du RDF et du RDFa, même si la discussion est ouverte à tous, certains acteurs ont le pouvoir de validation en dernière instance (Microsoft, Mozilla, Google et Apple pour les microdonnées, et Tim Berners-Lee pour le RDF et le RDFa). En comparant les syntaxes, nous avons également montré que tandis que certaines d'entre elles (RDF/RDFa) servent à sémantiser les informations après les avoir discrétisées, d'autres (les microformats) sont au service d'intérêts commerciaux visant à éditorialiser l'affichage des données sans forcément permettre (voire en empêchant) de relier les données entre elle. La dernière syntaxe

analysée (les microdonnées du Schema.org) vise à imposer un vocabulaire en proposant une syntaxe, et d'influencer les autres syntaxes en proposant un vocabulaire. Elle résulte autrement dit d'une stratégie prédatrice puisque du côté syntaxe (au niveau techno-applicatif) elle concurrence les microformats, et du côté vocabulaire (au niveau sémio-rhétorique) elle concurrence tous les autres vocabulaires existants. Parmi les créateurs de cette syntaxe se trouve Google, qui grâce à cette stratégie de prédation a pu relier les données à l'échelle du Web mais qui, étant donné ses intérêts financiers, au lieu de publier l'ontologie résultant de cette réunification, s'est contenté de proposer un outil d'exploration : le Knowledge Graph. Enfin, nous avons vu que tandis que toutes les syntaxes du code HTML suivent la langue naturelle, en balisant les mots et les expressions présents sur les pages Web, dans le corps (<body>) du document, l'usage du code Javascript consiste quant à lui à dissocier les données des informations, c'est-à-dire à dissocier l'espace consacré à la machine (techno-applicatif) de l'espace consacré au lecteur (sémio-rhétorique). Ces deux espaces deviennent détachables : le document, s'il voyage (par exemple si le contenu d'une page Web est recopié sur une autre page Web) ne voyagera plus avec ses balises sémantiques, ce qui obligera les acteurs intermédiaires à opérer une nouvelle opération de discrétisation/sémantisation.

Nous avons vu que dans le cas du RDF/RDFa le principe de *fragmentation*, propre au tropisme de *manipulabilité* donnait lieu à des possibilités de *génération* de nouveaux contenus d'autant plus nombreuses que plusieurs espaces de noms pouvaient être invoqués, alors que dans le cas du Schema.org un seul espace de nom est mis au service d'une fonction de fragmentation qui dès lors est plus simple, mais moins souple, et qui offre une plus grande manipulabilité au niveau techno-applicatif mais entraîne une diminution des possibles au niveau sémio-rhétorique. Quant aux microformats, nous avons vu que l'intention de ceux qui s'en saisissent est moins située au niveau sémio-rhétorique qu'au niveau techno-applicatif, alors même que l'intention de ceux qui les ont conçus est au contraire davantage située au niveau sémio-rhétorique qu'au niveau techno-applicatif. Nous pouvons émettre l'hypothèse qu'une telle ambivalence est liée au statut de « quasi-standard » des microformats.

En étudiant la syntaxe de ce qu'il nomme les « petites formes » générées par les API sous forme de widgets de code HTML préécrits, Samuel Goyet a récemment montré comment la discrétisation était mise « *au service d'une plus grande circulation des formes du texte* » (Goyet, 2017, p. 70) et comment elle pouvait faire l'objet d'une « optimisation » visant « standardiser » et à « industrialiser » la circulation des textes de réseau (*ibid.*, p. 90). En étudiant nous-mêmes la discrétisation, dans le cas du Web sémantique, nous avons montré que cette industrialisation pouvait concerner non pas seulement la forme du texte mais aussi son sens, ou plutôt son « faire sens ». Nous avons également montré que des stratégies de prédation pouvaient être mises en œuvre, étant donné notamment les enjeux économiques de ce « faire sens ». Il n'y a pas d'industrialisation sans standardisation, et il n'y a pas standardisation sans rapports de forces (Bush, 2011). Des visions s'affrontent, des projets, des intérêts. Une tectonique a lieu, dont dépendent les modalités concrètes de production et de mise en circulation à grande échelle des signes, c'est-à-dire des informations transformées en données qualifiables, puis qualifiées. Cette tectonique produira l'ordre ou le chaos, c'est selon, les sciences de l'information et de la communication étant à coup sûr les mieux placées pour comprendre à quoi et à qui tient ce « selon ».

Références bibliographiques

Bachimont, Bruno (1999), « De l'hypertexte à l'hypotexte : les parcours de la mémoire documentaire », *Technologies, Idéologies, Pratiques*, n° 4, 195-225.

Bachimont, Bruno (2010), *Le sens de la technique : le numérique et le calcul*, Paris : Les belles lettres, (collection « encre marine »).

Benezech, Danièle (1996), « La norme : une convention structurant les interactions technologiques et industrielles », *Revue d'Economie Industrielle*, n° 75, p. 27-44.

Bijker, Wiebe E. (1995), *Of Bicycles, Bakelites and Bulbs: Steps towards a Theory of Socio-technical Change*, Cambridge, MA : MIT Press.

Bouchardon Serge, Cailleau Isabelle, Crozat Stéphane, Bachimont Bruno, Hulin Thibaud (2011), « Explorer les possibles de l'écriture multimédia », *in* Paquienséguy Françoise (coord.), Dossier « Information publique : stratégies de production, dispositifs de diffusion et usages sociaux », *Les Enjeux de l'Information et de la Communication*, n° 12/2, 2011, p. 11-23, https://lesenjeux.univ-grenoble-alpes.fr/pageshtml/art2011.html#dossier

Bowker, Geoffrey C., Star, Suzan Leigh (1999), *Sorting Things Out: Classification and Its Consequences*, Cambridge, MA: MIT Press.

Busch, Lawrence (2011), *Standards: Recipe for Reality*, Cambridge, MA: MIT Press.

DeNardis, Laura (2009), *Protocol Politics: The Globalization of Internet Governance*, Cambridge, MA: MIT Press.

Crozat Stéphane, Bachimont Bruno, Cailleau Isabelle, Bouchardon Serge, Gaillard Ludovic (2011), « Éléments pour une théorie opérationnelle de l'écriture numérique », *Document numérique*, vol. 14, n°3, p. 9-33.

Ermoshina, Ksenia, Musiani, Francesca (2018), « Standardizing by Running Code: The Signal Protocol and De Facto Standardization in End-to-End Encrypted Messaging », *GigaNet: Global Internet Governance Academic Network, Symposium annuel 2017*.

Glen C. M. (2017), *Controlling Cyberspace. The politics of Internet Governance and Regulation*, Praeger.

Goyet, Samuel (2017), « Outils d'écriture du Web et industrie du texte: Du code informatique comme pratique lettrée », *Réseaux*, vol. 206, n° 6, p. 61-94.

Gruber, Thomas (1993), « A Translation Approach to Portable Ontology Specification », *Knowledge Acquisition*, vol. 5, p. 199-220.

Halpin, Harry (2017), « The Crisis of Standardizing DRM: The Case of W3C Encrypted Media Extensions », *SPACE 2017 - Seventh International Conference on Security, Privacy, and Applied Cryptography Engineering*, Décembre 2017, Goa, India, Springer, 10662, p. 10-29.

Khare, Rohit, Çelik, Tantek (2006), « Microformats: a pragmatic path to the semantic Web », *Proceedings of the 15th international conference on World Wide Web (WWW '06)*, ACM, New York, NY, p. 865-866.

Lelong Benoît, Mallard Alexandre (2000), « Présentation : la fabrication des normes », *Réseaux*, vol. 18, n°102, p. 9-34.

Lessig Lawrence (1999), *Code and Other Laws of Cyberspace*, New York, NY : Basic Books.

Malcolm Jeremy (2008), *Multi-Stakeholder governance and the Internet governance forum*, Australia: Terminus press.

Russel, Andrew L. (2003), « The W3C and its Patent Policy Controversy: A Case Study of Authority and Legitimacy in Internet Governance », *TPRC 2003*.

Sire Guillaume (2017), « Gouverner le HTML. Analyse du processus de normalisation du code HTML5 et de la controverse "Encrypted Media Extensions" », *Réseaux*, vol. 206, no. 6, p. 37-60.

Virili Francesco (2003), « Design, Sense-Making And Negotiation Activities In The" Web Services" Standardization Process », *MIS Quarterly Special Issue on Standard Making: A Critical Research Frontier for Information Systems*

Traces numériques et recherche scientifique au prisme du droit des données personnelles

Digital Traces and Scientific Research through the Lens of Data Protection Law

Huellas digitales e investigación científica a través del prisma de las leyes de protección de datos

Article inédit, mis en ligne le 15 novembre 2018.

Julien Rossi

Julien Rossi est doctorant au laboratoire COSTECH de l'Université de technologie de Compiègne. Sa thèse porte sur les politiques publiques de protection des données à caractère personnel, qu'il étudie en adoptant une approche communicationnelle de l'action publique permettant d'analyser le rôle des représentations politiques dans la formulation et la mise en œuvre de ces politiques publiques. julien.rossi@utc.fr

Jean-Edouard Bigot

Jean-Edouard Bigot est doctorant en sciences de l'information et de la communication au laboratoire COSTECH de l'Université de Technologie de Compiègne. Ses recherches portent sur les transformations numériques des méthodes de recherche en sciences sociales. Il développe en particulier une approche techno-sémiotique de l'instrumentation numérique en questionnant les enjeux épistémologiques et politiques associés à l'utilisation de dispositifs d'exploitation de données numériques dans l'étude des phénomènes socioculturels. jean-edouard.bigot@utc.fr

Plan de l'article

Résumé

La disponibilité croissante de masses de traces numériques inspire une nouvelle génération de projets de recherche numériquement instrumentés. Or l'exploitation de ces traces, en tant que le droit les qualifie de données à caractère personnel, entraîne l'application d'un corpus de normes juridiques dont le Règlement général de protection des données récemment adopté. Celui-ci prévoit un cadre dérogatoire pour la recherche scientifique, soutenu par une coalition d'acteurs issus de la recherche médicale, qui s'applique également aux sciences humaines et sociales (SHS). Le présent

article présente une synthèse de travaux de recherche sur ce cadre légal et la façon dont il vient interroger le rapport aux traces numériques dans la recherche en SHS.

Mots clés

Traces numériques, droit des données personnelles, sciences humaines et sociales, méthodes numériques.

Abstract

The ever-increasing masses of digital traces available online have inspired a new generation of digitally equipped research projects. As digital traces legally qualify as personal data, their use is subject to a set of rules such as the recently adopted General Data Protection Regulation. These regulations provide a special legal framework that is applicable to the use of personal data for scientific research purposes, it is supported by a coalition of actors involved in medical research but applies to humanities and social sciences as well. This article presents a summary of findings from studies on the way this legal framework interacts with digitally equipped research in humanities and social sciences and blurs the lines between digital traces and personal data.

Keywords

Digital traces, personal data protection law, humanities and social sciences, digital methods.

Resumen

La disponibilidad creciente de masas de huellas digitales en línea inspira una nueva generación de proyectos de investigación optimizados para trabajo digital. Las leyes de protección de datos personales, entre las cuales se haya el reciente Reglamento general de protección de datos, incluyen esas huellas digitales en la categoría jurídica de los datos personales. Sin embargo, el Reglamento incluye un régimen derogatorio para el empleo de datos personales con fines de investigación científica, que fue apoyada por una coalición procedente de la investigación médica. Este articulo presenta de manera resumida las conclusiones de investigaciones sobre este marco legal, que se aplica también a las ciencias humanes y sociales, y que contribuye a difuminar el límite entre huellas digitales y datos personales.

Palabras clave

Huellas digitales, ley de protección de datos personales, ciencias humanas y sociales, métodos digitales.

Introduction

Le numérique prend progressivement une place accrue dans le champ des sciences humaines et sociales (SHS) (Bourdeloie, 2014) qui se trouverait ainsi « mis au défi » (Diminescu et Wieviorka,

2015) au point que certains observateurs proclament l'avènement d'une « troisième génération de SHS » (Boullier, 2015 A et B). Dans ce contexte, « s'équiper numériquement »[1] deviendrait incontournable, même pour des disciplines *a priori* éloignées du numérique. Ces projets ont notamment pour matériau les « traces numériques ». Or qu'il s'agisse classiquement d'entretiens, de réponses à des questionnaires, ou de traces numériques, ces données entrent toutes dans la catégorie juridique des données à caractère personnel, entraînant l'application à ces projets du Règlement général de protection des données[2] (RGPD) adopté en 2016 et entrant en vigueur le 25 mai 2018, venu mettre à jour sans pour autant en remettre en cause les grands principes les lois antérieures telles que la loi française Informatique et Libertés de 1978 (Gellert, 2016).

Comment, si elle a lieu, s'établit la médiation entre ce droit et les projets de recherche évoqués plus haut, consommateurs de « traces numériques » ? Quelle est sa matérialité sur le terrain ?

Pour répondre à ces questions nous avons conduit une recherche en quatre principales étapes. En parcourant la littérature académique sur les projets « numériquement équipés » en SHS et en portant un regard réflexif sur nos propres pratiques d'exploitation de données numériques, nous avons cherché à déterminer le régime juridique applicable en théorie à ce type de recherches, ainsi que la généalogie de ce régime et le ou les référentiels[3] sous-jacents en ayant légitimé l'adoption. Nous avons étudié les échanges lors de deux journées d'études organisées sur l'application à la recherche en sciences sociales du droit des données personnelles[4]. Ces journées visaient à faire dialoguer des professionnels de la protection des données et des enseignants-chercheurs. Elles ont formé des arènes publiques éphémères permettant l'expression de débats (Badouard, Mabi, et Monnoyer-Smith, 2016) et de controverses (Smadja, 2012) n'ayant pas la place d'être exposés ici en détail, mais cependant intégrés à la présente analyse. Enfin, nous avons mené une campagne d'entretiens avec les chercheurs et chercheuses d'un projet de recherche financé par l'ANR, comprenant notamment des chercheurs en SIC et ayant impliqué le traitement statistique de données numériques personnelles dont des traces collectées en ligne.

Les données que nous avons collectées ont été analysées sous l'angle d'une approche communicationnelle du droit et des politiques publiques. Cette approche nous invite à considérer que les normes juridiques sont des instruments (Lascoumes, 2004) dont le contenu est le fruit d'un processus délibératif impliquant des controverses parmi des acteurs impliqués dans un même champ de l'action publique (Dubois, 2014) ou sous-système de politique publique (Sabatier, 1998 ; Bergeron, Surel, et Valluy, 1998). Ces acteurs sont regroupés au sein de coalitions qui visent à imposer leur référentiel politique à l'ensemble du champ pour légitimer les dispositions juridiques concrètes pour lesquelles ils militent. La cartographie de ces coalitions permet de rendre compte de la structuration des controverses politiques en partant de l'hypothèse qu'elles se forment autour de

. . . . . .

1 Nous nommons « recherche numériquement équipée » toute recherche s'appuyant sur des dispositifs d'exploitation de données numériques intervenant en position d'instruments dans les processus de connaissance (Bigot, 2018).

2 Règlement 2016/679/UE du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE.

3 Nous empruntons la notion de référentiel à Pierre Müller pour qui : « Elaborer une politique publique consiste donc d'abord à construire une représentation, une image de la réalité sur laquelle on veut intervenir. C'est en référence à cette image cognitive que les acteurs organisent leur perception du problème, confrontent leurs solutions et définissent leurs propositions d'action : cette vision du monde est le référentiel d'une politique ». (Müller, 2011, p. 57)

4 Il s'agit de la journée « Données personnelles en milieu universitaire : quelles questions se poser ? » organisée à l'Institut des sciences de la communication le 12 janvier 2017 et de la journée « Données personnelles et sensibles : quels enjeux pour la recherche en SHS ? » organisée par l'Institut des sciences sociales du politique à l'Université Paris Nanterre le 7 novembre 2017.

référentiels idéologiques communs. Les textes juridiques rédigés puis adoptés par ces acteurs sont des outils techniques de gouvernementalité dont le fonctionnement repose en effet sur le principe de la performativité. Ceci les soumet à des conditions de succès, parmi lesquelles le fait que « les participants aient l'intention d'adopter le comportement impliqué » par l'invocation d'un énoncé performatif qui lui-même suppose « certaines pensées ou sentiments » (Austin, 1970, p. 49) permettant d'en saisir le sens (voir également : Reinach, 2004). Or un texte de loi publié est un support matériel de médiation dont le contenu fait l'objet d'une réinterprétation par son public, et de controverses autour son interprétation, débattue dans des procès. Les rédacteurs d'un texte de loi peuvent d'ailleurs laisser à dessein des ambivalences rendre compatibles des visions incompatibles lorsqu'un compromis est nécessaire (voir sur un sujet similaire : Krieg-Planque, 2010), renvoyant en pratique au pouvoir judiciaire le soin de trancher. La réinterprétation du texte conditionne le sens donné à sa mise en actes, et donc les usages potentiels du texte juridique en tant qu'outil politique. On ne peut supposer d'équivalence entre le sens donné à un texte juridique par son rédacteur et celui que lui donne son lecteur ou son public, ce dernier pouvant ne prendre connaissance de son contenu qu'après une série de médiations successives.

Dans notre cas d'étude, ces textes juridiques sont réinterprétés en fonction de référentiels propres, non pas tant aux convictions individuelles des individus que nous avons interrogés, mais au référentiel sectoriel du champ de la recherche scientifique qui a ses propres normes, discours, valeurs et logiques d'intérêts. Nous verrons d'ailleurs que ce référentiel scientifique et son ethos peuvent parfois entrer en conflit avec les convictions personnelles des chercheurs ayant participé à l'enquête.

Promesses et critiques des « traces numériques »

L'avènement d'une recherche numériquement équipée en SHS s'accompagne de discours, plus ou moins utopistes, concernant en particulier un renouvellement « positif », au double sens de « bénéfique » et de « positiviste », des paradigmes (Kuhn, 1962) des disciplines des sciences de la culture grâce à la possibilité offerte par les technologies numériques de se saisir de données sociologiques massivement récoltées sur les réseaux informatiques, envisagées comme des traces de pratiques sociales. Ces discours et les promesses dont ils sont porteurs sont particulièrement présents dans le projet du courant des dites méthodes numériques (Rogers, 2009 et 2013 ; Venturini et Latour, 2009 ; Rieder, 2010 ; Rieder et Röhle, 2012 ; Venturini et al., 2014).

Selon le constat fréquemment émis par les promoteurs des méthodes numériques, le développement des technologies numériques s'accompagne d'une production massive de données censées refléter les échanges sociaux qui se produisent en ligne, et de plus en plus ceux qui ont lieu hors ligne aussi. Ces données numériques massives proviendraient d'abord des données d'utilisation des plateformes de réseautage social comme Facebook ou Twitter mais aussi, plus globalement, de l'usage de n'importe quelle application informatique. Il est commun de reconnaître que l'utilisation des médias informatisés s'accompagne toujours de la production et du stockage de deux types d'informations : d'une part les informations produites par les utilisateurs (création et partage de contenus, informations personnelles délivrées lors de la création de profils, etc.) et d'autre part les informations résultant d'un enregistrement automatique des actions effectuées par les utilisateurs (temps passé sur un site, contenus consultés, les liens hypertextes actionnés, etc.). C'est ce dernier type de données, distinctes par le caractère non-intentionnel de leur production, que les praticiens des méthodes numériques désignent généralement par le terme de « traces » (voir : Venturini, 2012).

Toute interaction sociale médiée par une application numérique tendrait donc aujourd'hui à être tracée et cette nouvelle traçabilité sociale serait rendue sous la forme d'un vaste ensemble de

données numériques largement accessibles sur les réseaux informatiques publics, via les API des sites web et via des services, publics ou privés, dédiés à la diffusion d'informations numérisées. Les technologies numériques offriraient ainsi un réservoir potentiellement illimité de données sociales dont il serait aisé de se saisir dans un contexte de recherche, celles-ci représentant une opportunité pour les SHS de mieux comprendre les transformations contemporaines de la société.

La disponibilité de ces traces numériques permettrait un accès facilité au terrain pour les chercheurs grâce à la fois à une réduction des « coûts » humains, économiques et temporels de la constitution de corpus, et une ouverture croissante des données qui les rendent plus aisées à obtenir, à agréger et à manipuler. En découlerait l'autorisation d'une observation *in vivo* des pratiques sociales qui se produisent sur le Web :

> « *L'intérêt des médias électroniques est que toute interaction qui les traverse laisse des traces qui peuvent être facilement enregistrées, massivement stockées et aisément récupérées [...] offrant aux sciences sociales plus de données qu'elles n'en ont jamais rêvées. [...] Nées dans une époque de pénurie, les sciences sociales entrent dans un âge d'abondance* ». (Venturini et Latour, 2009)

La possibilité d'une transformation des sciences sociales par l'exploitation appareillée de traces numériques, aux fondements du projet des méthodes numériques, a fait l'objet d'importantes critiques émanant en particulier des SIC. Elle sont relatives à la notion de « trace », à son épistémologie et aux enjeux de pouvoir qui la traversent (Merzeau 2009 et 2013 ; Flon et al., 2009 ; Jeanneret, 2011 et 2013 ; Collomb, 2016). Ces critiques visent fondamentalement à déconstruire l'apparente évidence de la trace et s'opposent à une conception naturalisante des traces numériques. Yves Jeanneret invite ainsi à dépasser une conception indicielle des traces numériques en appelant à réinscrire ces objets dans les logiques communicationnelles complexes qui les sous-tendent. Dans cette perspective communicationnelle, les traces sont des objets sémiotiques qui résultent d'une construction culturelle relevant d'une élaboration documentaire : documentation, médiatisation, archivage, etc. Parce qu'elles sont toujours liées à des pratiques d'écriture, les traces numériques doivent être envisagées comme des inscriptions, qu'elles soient intentionnelles ou non. Si ces données sont générées et collectées de manière automatique, elles sont aussi traitées, c'est-à-dire matérialisées dans un espace qui leur donne sens, les organise, leur assigne une place, et les inscrit dans un projet interprétatif. Il y a donc un problème à considérer les données récoltées sur les réseaux comme des traces de pratiques sociales, sans en interroger les conditions de traçabilité et de traitement. Aussi, il convient d'adopter une définition de la trace qui prenne en compte les médiations à la fois logistiques et sémiotiques qui permettent d'interpréter certaines productions communicationnelles comme traces d'identité, traces d'usages ou traces de phénomènes sociaux (Jeanneret, 2011). De la même manière, Cléo Collomb (2016) insiste sur le fait que ce que l'on désigne communément comme des traces doive d'abord être considéré comme un ensemble d'inscriptions réalisées par l'interaction d'un agent humain et d'un programme informatique qui double, selon une logique de supplément et selon des modalités spécifiques à la médiation techno-sémiotique qu'il opère, des activités d'utilisation dans un média informatisé. Nous souscrivons ici à cet ensemble de définitions apporté par les SIC, centré sur la dimension à la fois matérielle, technique, et sémiotique des traces numériques.

Les réflexions critiques conduites par les SIC au sujet des traces numériques et, plus largement, sur les questions liées aux imaginaires du « *big data* » et leur rapport avec la problématique de la surveillance par les données (Carré et Panico, 2011 ; Rouvroy et Berns, 2013) apportent des éclairages originaux et précieux aux enjeux des données numériques. En outre, il nous apparaît tout aussi important de pouvoir aborder les problématiques associées au cadre spécifique des données à caractère personnel collectées à des fins de recherche, et la façon dont celui-ci est reçu par les chercheurs qui font usage de données numériques.

Le droit applicable et sa généalogie

En prenant appui sur les études précédemment menées sur la genèse du droit à la protection des données (Hondius, 1975 ; Newman, 2008 ; Vitalis, 2009 ; Atten, 2013 ; Gonzalez Fuster, 2014) nous avons constitué un corpus de documents d'archives issus principalement des groupes de travail du Conseil de l'Europe, de l'OCDE puis de l'Union européenne, à l'origine des notions juridiques de « donnée à caractère personnel » et de « protection des données ». Ces documents ont été complétés par des entretiens qualitatifs avec 16 décideurs publics, pour certains d'entre eux depuis la fin des années 1960, impliqués dans le champ de l'action publique relative à la protection des données à caractère personnel. Ceux-ci nous ont permis de conclure à l'existence d'un réseau d'acteurs réunis dans une coalition de cause reposant sur le « paradigme de la vie privée » (Bennett et Raab, 2003) inspiré de la philosophie utilitariste de John Stuart Mill (1989) et des travaux de Michel Foucault sur le panoptique (Foucault, 1975 ; Simon, 2002). Pour ces acteurs, l'informatisation de la société aboutit à la mise en place de dispositifs de surveillance panoptiques mettant en péril la liberté et l'autonomie des individus notamment par la mise en place d'une gouvernementalité par le calcul (Rouvroy et Berns, 2013) et par la sophistication de mécanismes de discrimination fondée sur l'exploitation de données personnelles (Lyon, 2015). Cette conception est devenue le référentiel sectoriel dominant dans le sous-système de politique publique de la protection des données. Elle suppose de laisser l'individu maître de décider du sort de ses données personnelles par un mécanisme de « droit à l'autodétermination informationnelle » (Tribunal constitutionnel fédéral allemand, 1983) plutôt que de définir par avance et pour tout le monde ce qui relève de la sphère privée ou de la sphère publique.

Cette logique de droit à l'auto-détermination informationnelle, consacrée par le législateur européen, n'opère pas de choix a priori entre ce qui relève du privé et du public. La définition juridique qui s'est stabilisée progressivement dans différents textes juridiques à partir des années 1970 laisse ainsi à l'individu le pouvoir de négocier la frontière entre public et privé :

> « *Toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée «personne concernée»); est réputée être une ″personne physique identifiable″ une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale* ». (art. 4 §1 du RGPD)

Comme il a été plusieurs fois rappelé par la Cour de justice de l'Union européenne :

> « *[...] les notions de ″données à caractère personnel″ [...] et de ″données relatives à la vie privée″ ne se confondent pas [...]* ». (CJUE 2015, pt. 32)

De surcroît, « il n'est pas requis que toutes les informations permettant d'identifier la personne concernée se trouvent entre les mains d'une seule personne » (CJUE 2017, pt. 31) pour qu'une donnée soit qualifiée de personnelle. De nombreux travaux en informatique repris par la doctrine juridique ont en effet montré la difficulté pratique d'anonymiser des données personnelles. Il demeure généralement possible, par croisement de données quasi-identifiantes, de retrouver une personne, même indirectement (Sweeney, 2000 ; G29, 2007 ; Ohm, 2010 ; Mascetti et al, 2013 ; G29, 2014 ; de Montjoie et al, 2015 ; Rossi, 2015), ce qui est suffisant pour entrer dans le cadre de la définition citée ci-dessus. En outre, comme le rappelle le considérant 26 du RGPD au sujet des données pseudonymes :

> « *[...]Les données à caractère personnel qui ont fait l'objet d'une pseudonymisation et qui pourraient être attribuées à une personne physique par le recours à des informations*

> *supplémentaires devraient être considérées comme des informations concernant une personne physique identifiable [...] »*. (RGPD, cons. 26)

En pratique, toute trace numérique devient alors susceptible d'être qualifiée de donnée à caractère personnel, entraînant l'application d'un grand nombre de contraintes pouvant comprendre le consentement de la personne concernée, la limitation de la durée de conservation des données ou encore une obligation d'information complète et transparente qui peuvent poser des problèmes méthodologiques :

> « *À mon sens, quand je vous écoute, je me demande si tout un tas d'enquêtes sociologiques, notamment, qui ont été menées je pense, sur les milieux soit très difficiles d'accès, soit ésotériques, auraient été possibles avec ce type d'impératifs légaux* ». (ISP, 2017b, 11m 10s secondes à 11m 32s)

> « *Si on utilisait un dispositif formel d'interrogation d'entretiens, on n'allait pas potentiellement […] avoir le même type de parole que celle qui nous est dite de manière informelle* ». (entretien 8)

De plus, il est parfois impossible de recueillir le consentement des personnes ou de les informer de façon parfaitement transparente. Cela peut même s'avérer dangereux pour les chercheurs notamment lorsqu'ils travaillent sur des groupes sociaux violents, comme en a témoigné Fanny Georges (2018), responsable scientifique du projet MINWEB sur la circulation de messages politiques dans la zone de conflits qu'est le Sahara.

L'article 89 du RGPD fournit à la recherche scientifique un certain nombre de dérogations pour répondre à ces préoccupations, dont la portée exacte doit cependant être précisée par les législateurs nationaux. L'appréciation du caractère d'intérêt public de la recherche menée sera un élément d'appréciation pour la CNIL pour déterminer dans quelle mesure ces dérogations pourront s'appliquer :

> « *Par exemple, […] pour des recherches qui seraient considérées comme d'intérêt public, dans le cadre d'une mission de service public, le consentement peut ne pas être recueilli. [...] Même sur le plan de l'information […] on peut aussi considérer que l'information des personnes représenterait un effort disproportionné par rapport à l'intérêt même de la recherche qui est menée et de la finalité poursuivie* ». (Sophie Genvresse, ISP, 2017b, 12m01s à 12m37s)

Or la légitimité de ces dérogations a fait l'objet de remises en cause au cours du processus d'adoption du RGPD.

L'article 89 n'a fait l'objet que de peu de lobbying au cours du processus d'adoption du RGPD. Ainsi, sur 227 *position papers* en anglais recensés dans notre corpus, seuls 21, soit à peine 10 %, portaient au moins en partie sur le statut de la recherche scientifique. Ces 21 documents ont été rédigés par 15 organismes répartis comme décrit par les graphiques ci-dessous, par secteur d'activité et par type :

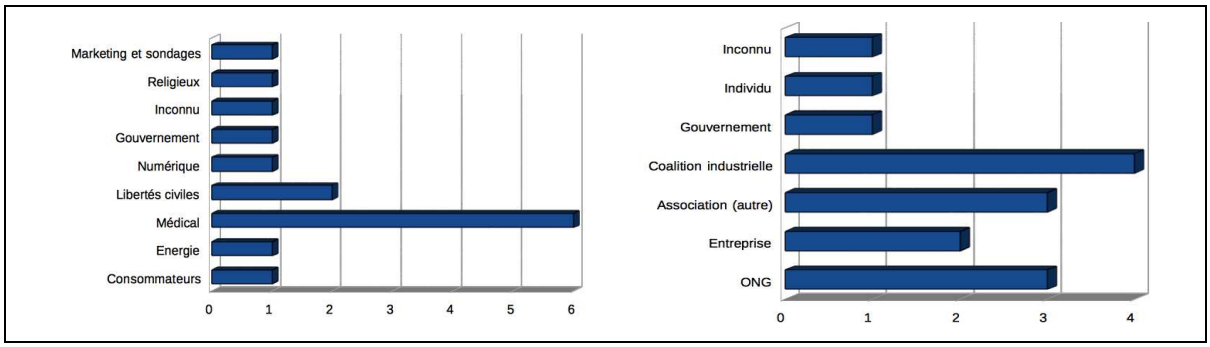


Figure 1. Répartition des organismes par domaine d'activité (à gauche) et par type (à droite).

Ces organismes se regroupent autour de deux coalitions de cause.

La coalition médico-pharmaceutique, avec notamment la Fédération européenne des académies de médecine (FEAM) et Comité de coordination européen de l'industrie radiologique, électromédicale et d'informatique médicale (COCIR), portait une position communautariste au sens d'Amitai Etzioni (1999). En ce sens, ils privilégient l'intérêt de la communauté et les objectifs de santé publique aux intérêts individuels de vie privée, ainsi qu'une confiance dans l'utilité du traitement massif de données de santé passant par la réutilisation des dossiers de patients pour la recherche médicale. Cette position justifiait selon eux des dérogations importantes pour la recherche, comme l'abaissement des exigences d'anonymisation au profit d'une simple pseudonymisation et une limitation considérable des droits des enquêtés quant à leurs données personnelles dès lors qu'elles étaient exploitées, avec ou sans leur consentement, dans une finalité de recherche.

L'autre coalition, portée par les associations Bits of Freedom et European Digital Rights, visait à exclure certaines données sensibles et celles portant sur des mineurs, sauf consentement des enquêtés, selon le cadre dérogatoire de l'article 89. Son périmètre aurait aussi été restreint à des recherches perçues comme plus légitimes que d'autres, par exemple celles bénéficiant de la protection de l'article 13 de la Charte des droits fondamentaux de l'Union européenne, qui protège la liberté académique. Cette coalition adhère au référentiel du paradigme libéral de la vie privée (Bennett er Raab, 2003) précédemment évoqué.

Ainsi pour l'association European Digital Rights :

> « *Processing of sensitive data for historical, statistical and scientific research purposes is not as urgent or compelling as public health or social protection. Consequently, there is no need to introduce an exception, based on national law, which would put them on the same level as the other listed justifications, which risks undermining fundamental rights, legal certainty and the single market*[5] ». (EDRi, 2012 p. 19)

Le copier-coller de ce texte se retrouve dans le projet de rapport de Jan-Philip Albrecht, rapporteur du projet de règlement qui a donné le RGPD (Albrecht, 2013, p. 24), qui voit en les promesses du Big Data un avenir dystopique :

> « *If nothing else, would it not make sense to ignore or even oppose the Big Data efficiency enhancement logic simply because it makes us into machines being driven to the limits of their efficiency. For that logic will make society into a sterile place full of clean clones*

. . . . . .

5 « Le traitement de données sensibles à des fins historiques, statistiques ou scientifiques n'est ni aussi urgent ni aussi impératif que la santé publique ou la protection sociale. Par conséquent, il n'y a aucun besoin d'introduire une exception, fondée sur le droit national, qui mettrait ces justifications au même niveau, car cela porterait atteinte aux libertés fondamentales, au principe de sécurité juridique et au marché unique ». (Traduction des auteurs)

> *[…]. The mass analysis of our personal data makes the neoliberal logic of the exploitation of human life into the all-encompassing mantra of the post-modern world*[6] ». (Albrecht, 2015, p. 93)

A défaut de participation au débat par des chercheurs en SHS sur ce qu'est devenu l'article 89 du RGPD dans les arènes institutionnelles pertinentes pour viser à en fonder la légitimité sur d'autres bases que les attentes du *Big Data* pour la recherche médicale et pharmaceutique, cette reprise n'a rien de surprenant.

Comme nous l'avons vu, le cadre juridique du RGPD, qui n'est qu'une évolution de règles en vigueur en France depuis la loi Informatique et Libertés de 1978, n'est pourtant pas sans poser des difficultés, notamment méthodologiques, aux SHS. Nous avons donc mené une enquête sur un projet de recherche concret pour étudier la réception de ces contraintes juridiques faite sur le terrain par des chercheurs en SHS, dont une forte proportion est située en SIC.

La protection des données personnelles sur le terrain d'un projet en SHS

Parmi les chercheurs que nous avons interrogés dans le cadre d'entretiens semi-directifs, huit ont participé ensemble à un projet de recherche en sociologie et sciences de l'information et de la communication financé par l'ANR[7]. Les questions posées visaient à comprendre leur réception du droit des données personnelles, en lien avec les questionnements éthiques devant lesquels leurs collectes de traces et données numériques les plaçaient.

Le projet de recherche dans lequel les répondants étaient investis collectait des données personnelles de plusieurs façons. D'une part, il s'agissait d'effectuer une collecte de données en ligne numériquement instrumentée, notamment par le biais des réseaux socio-numériques. D'autre part, des données étaient générées par des questionnaires et des entretiens retranscrits avec des enquêtés dont la caractéristique commune était d'avoir vécu le type de traumatisme personnel sur lequel portait le projet. L'ensemble de ces données fit l'objet d'un traitement statistique.

L'absence presque complète de médiation entre les chercheurs interrogés et les acteurs de la protection des données à caractère personnel, qu'il s'agisse d'associations ou de juristes externes à l'université tels que ceux de la CNIL, ou qu'il s'agisse de services juridiques internes et notamment des correspondants informatique et libertés (CIL), est un constat qui s'est très vite imposé dans tous les entretiens.

Les CIL sont des personnes chargées de superviser la mise en conformité aux règles de la protection des données personnelles de leur employeur (Rossi, 2015). Appelées « délégués à la protection des données » dans le RGPD, leur désignation était encouragée depuis 2004 par la loi Informatique et Libertés. En juin 2017, grâce aux jeux de données rendus disponibles sur data.gouv.fr, nous avions ainsi répertorié 137 CIL désignés par des établissements publics d'enseignement supérieur et de recherche, tous internes à l'organisation. Parmi ceux-ci, seuls 24 % disposaient d'un site web permettant d'identifier les coordonnées du CIL, et 20% n'affichaient même pas les mentions légales

. . . . . .

6 « Cela n'aurait-il pas de sens d'ignorer voire de s'opposer à la logique d'accroissement de l'efficacité propre au Big Data, pour la simple raison que cela fait de nouveau des machines poussées à la limite de leur efficacité ? Car cette logique transformera la société en un lieu stérile plein de clones propres [...]. L'analyse de masse de nos données personnelles transforme la logique néolibérale d'exploitation de la vie humaine en un mantra qui englobe l'ensemble de notre monde post-moderne ». (Traduction des auteurs)

7 Nous tenons à remercier chaleureusement chacune de ces personnes pour leurs contributions. Pour des raisons de confidentialité, le nom du projet et, sur leur demande, les noms de certains participants à l'enquête ont été anonymisés.

obligatoires. Cela est symptomatique d'une faiblesse institutionnelle des CIL, qui n'ont pas forcément les moyens d'exercer leurs missions.

Ainsi même, les enquêtés qui avaient connaissance de l'existence du CIL nous ont indiqué n'avoir jamais reçu de réponse à leurs sollicitations de la part de celui ou celle de leur université :

> « *Au démarrage, [...] on a pris contact avec le CIL de l'Université XXX. Pour le dire très simplement, on n'a jamais eu de retour de l'Université XXX* ». (entretien 2)

L'absence de communication entre le champ de la protection des données et celui de la recherche en SHS à l'université contribue à expliquer des approximations juridiques et des confusions exprimées par les membres du projet interrogés, et ce nonobstant l'expression de leurs préoccupations éthiques vis-à-vis des enquêtés et leurs données collectées.

Par exemple, ainsi que l'exprimait un CIL présent lors de la journée d'étude du 12 janvier 2017 :

> « *Il y a souvent la confusion entre un projet de recherche totalement anonyme [...] et le rendu anonyme, en fait. Au départ on traite quand même bien des données personnelles qui n'ont rien d'anonyme. C'est juste le résultat qui le devient* ».

Son constat d'une confusion entre l'anonymat des données et celui du rendu est recoupé par les entretiens que nous avons réalisés :

> « *J'estimais [que] le côté problématique, ce n'était pas tellement les bases dans les serveurs. [...] Ce qui me posait problème, c'était plutôt au niveau de la publication. C'est-à-dire ce que j'allais pouvoir raconter sur des gens qui n'avaient rien demandé* ». (entretien 6)

> « *Je ne me suis pas posé la question des données informatiques : qu'est-ce qui restait des fichiers par exemple ? Je ne sais même pas où est-ce qu'ils sont ces fichiers aujourd'hui* ». (entretien 8)

Il existe par ailleurs une confusion fréquente entre la notion juridique de « donnée à caractère personnel » et celle de donnée relative à l'intimité des personnes, accompagnée du sentiment que travailler sur des données accessibles publiquement permet de contourner l'application de la loi alors que cet élément n'est qu'un élément à prendre en compte parmi d'autres dans l'analyse juridique :

> « *Les personnes ont mis leurs données sur internet. Donc elles sont censées savoir qu'elles les ont rendues publiques [...]. Et c'est là aussi où je pense que la question juridique n'est pas suffisante pour moi. [...] Oui, d'un point de vue juridique on peut considérer qu'elles sont publiques* » (entretien 3).

Si la réponse ne leur semblait pouvoir venir d'un usage du droit des données à caractère personnel, les chercheurs interrogés ont tous fait part de réflexions éthiques. Ces approches, parfois divergentes et sujettes à controverse au sein de l'équipe, ont pour point commun cette importance accordée à la préservation de l'anonymat des enquêtés dans les résultats publiés tant pour des raisons méthodologiques visant à permettre aux enquêtés une expression plus libre et donc de meilleure qualité, que pour des raisons éthiques.

Confirmant les travaux de Latanya Sweeney (2000) et Paul Ohm (2010) sur la difficulté d'anonymiser de façon robuste, c'est-à-dire sans risque aucun de ré-idendification des données de recherche par qui que ce soit, les chercheurs interrogés ont fait part de réflexions élaborées sur la façon d'aboutir à l'anonymat dans la publication. L'usage d'un corpus documentaire issu d'extractions de données du Web pose un problème spécifique et nouveau renforçant la difficulté de l'anonymisation :

> « *Même si j'ai anonymisé tous les noms de personnes [...] j'ai recopié sur mon [document public] des messages tels quels, tels qu'ils sont publiés sur le site internet. C'est-à-dire qu'une*

> *personne qui veut retrouver l'identité de la personne qui a écrit, il suffit de taper sur Google le message [...] Donc en fait je n'ai pas du tout anonymisé* ». (entretien 3)

La spécificité de l'enquête à partir de données publiées sur le Web est en effet la simplicité de la ré-identification par croisement avec des données publiquement accessibles :

> « *Si ça avait été fait dans le cadre d'un entretien privé, personne n'aurait pu remonter à l'information. Là [sur internet] tout le monde peut remonter à l'information* ». (entretien 3)

Cette question se pose également dans le cas de la collecte par questionnaires. Le questionnaire diffusé par la Société française des sciences de l'information et de la communication (SFSIC) sur l'éthique de la recherche en ligne est symptomatique de la difficulté rencontrée. Outre divers manquements à la législation tel que l'absence des mentions légales obligatoires, alors qu'il prétendait être anonyme, et qu'il le sera sans doute dans le rendu qui en sera effectué, il contenait des injonctions à la fourniture, sous peine d'être exclu de l'enquête, de renseignements individuels précis. Ces derniers pouvaient aboutir à la possibilité d'identifier une bonne partie des répondants par croisement avec des informations accessibles sur le Web comme la combinaison des langues parlées par la personne (en regardant les langues des publications de la personne, ou son CV en ligne), ou encore leur statut et leur date de naissance notamment lorsqu'il s'agit de dirigeants d'institutions académiques. Ceci était d'autant plus problématique que le questionnaire incitait les répondants à formuler des critiques vis-à-vis d'institutions et de lieux de pouvoir dont la SFSIC elle-même.

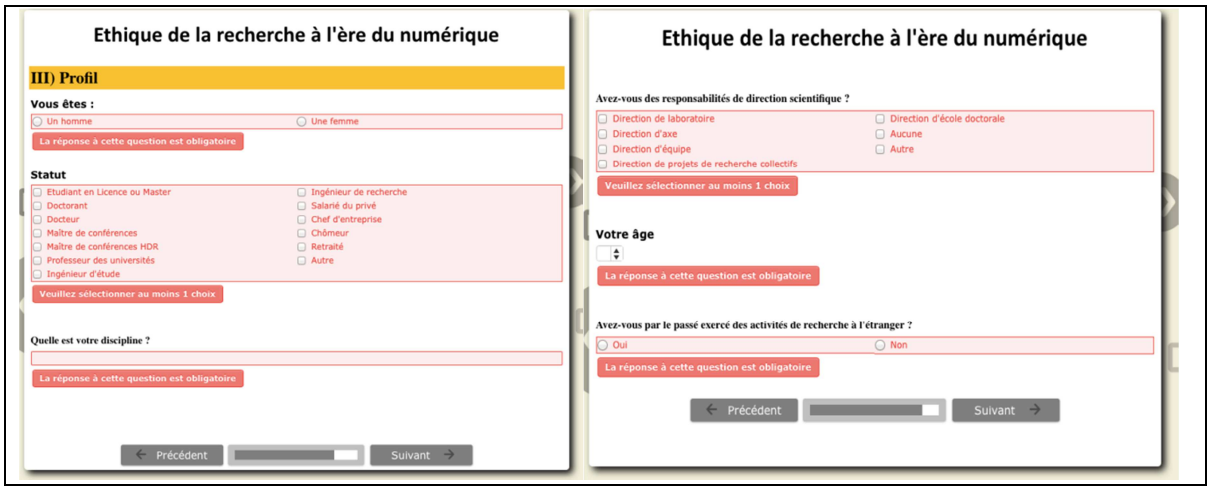


Figure 2. Captures d'écran du questionnaire de la SFSIC (2017)

Pour ne pas simplifier les choses, il peut arriver que l'anonymat, y compris lors de la publication, ne soit impossible ou qu'il entre en contradiction avec les finalités d'une recherche perçue comme ayant un intérêt public. Cela est typiquement le cas lorsqu'un chercheur travaille sur une personnalité publique.

Il peut arriver que des pressions s'exercent par ailleurs sur les chercheurs pour qu'ils partagent des données de recherche pourtant confidentielles. Ces pressions peuvent venir des services de police ou de renseignement dans certains cas, de l'industrie, mais également d'institutions de recherche incitant au partage et à l'ouverture des données pour favoriser la transparence de la recherche et le partage des connaissances :

> « *J'étais forcée pour mon DESS de publier [...] Et j'avais fait une erreur. Parce qu'en fait, mon DESS me forçait de partager un [rapport contenant des données confidentielles]* ». (entretien 7)

Cette pression avait également été évoquée par une des conférencières de la journée d'étude du 12 janvier 2017 qui y présenta la façon dont elle avait mobilisé des arguments tirés du droit à la

protection des données pour résister à des pressions visant à ce qu'elle abaisse le niveau d'anonymisation de ses résultats d'enquête sur des données au caractère politique sensible :

> « *Il s'en est suivi une négociation, au cours de laquelle j'ai à nouveau été confrontée à ma question initiale : pour qui ? Notamment, en en discutant au sein de l'institut polonais, j'ai entendu qu'on ne fait pas [...] de recherche pour les [enquêtés]. Il s'agissait donc de respecter à la fois la loi et les normes du milieu, mais aussi de faire un choix de nature plus morale, de se prononcer sur la question de responsabilité et de peser les bénéfices d'un large accès au savoir contre le risque de nuire aux participants et de trahir leur confiance* ». (Skowronska, 2017)

Certains chercheurs débutants ont exprimé ressentir un décalage entre leurs convictions personnelles et un certain *ethos* du partage et de la transparence de la recherche scientifique, qui ne les amène cependant pas, dans le cas du projet ANR que nous avons étudié, à se tourner vers le droit. De façon générale l'analyse des entretiens a montré que le lien entre les convictions individuelles exprimées par les enquêtés et l'usage des outils juridiques de protection des données était généralement faible, même lorsqu'ils avaient conscience par ailleurs de l'existence de ce droit. Les cas où la notion de confidentialité des « données personnelles » est évoqué sous un angle juridique par les enquêtés relevaient en réalité souvent d'une confidentialité au titre du droit de la propriété intellectuelle et visait à se protéger d'une dépossession de ses résultats de recherche face à l'institution universitaire ou à des collègues en qui la confiance n'est pas pleinement assurée.

Cette absence relative de questionnement en sciences humaines et sociales sur le cadre juridique de la collecte de « traces numériques », qui n'empêche en rien le déroulement de réflexions éthiques individuelles, provenait dans certains cas d'un sentiment initial de dépersonnalisation de ces traces et d'impression initiale d'anonymat erronée ou mal appréciée. L'apparition de l'aspect personnel des traces, dès lors qu'elles sont structurées sous une forme en permettant l'analyse, peut cependant prendre le chercheur par surprise :

> « *J'essayais de faire une variable qui pouvait coder [une caractéristique du comportement des personnes sur lesquelles les données étaient recueillies]. [...] En gros j'attribuais un score à chaque modalité. Plus on allait sur des plateformes d'un certain type, plus on avait un score élevé, en gros. Et c'est vrai que moi j'ai sorti ça, juste en statistique. Et après en fait ça nous donnait [une variable de score comportemental]. Enfin c'est hyper bizarre. Enfin y a des choses comme ça où on le fait statistiquement mais avec notre objet ça colle pas. Et là-dessus, même si c'était pertinent, on a complètement laissé tomber parce que c'était très bizarre en fait. [...] Là ce n'était vraiment pas une question statistique, c'était vraiment une question éthique je pense* ». (entretien 5)

Pour finir, même lorsque des données sont générées hors ligne par exemple par la sollicitation d'entretiens, leur analyse suppose aujourd'hui une inscription sur un support numérique, souvent en ligne, qui peut aboutir à une perte de contrôle informationnel sur les données. L'usage de services d'hébergement en ligne des données que les universités, dans le cas que nous avons étudié, refusaient de fournir elles-mêmes, renforçait chez les chercheurs ayant répondu à notre enquête un sentiment d'impuissance quant à leur capacité à sécuriser l'accès aux données qu'ils recueillent et à véritablement garantir le contrôle de leur circulation.

Conclusion

Les entretiens que nous avons réalisés montrent une absence de médiation entre les chercheurs impliqués dans le projet étudié et le champ de la protection des données à caractère personnel, en

dehors de quelques rares occasions comme les deux journées d'étude évoquées et répertoriées en bibliographie. Dès lors, une multitude de pratiques, reposant essentiellement sur l'anonymisation des publications, sont développées pour préserver la confidentialité des enquêtés. Le passage au numérique affaiblit cependant la frontière entre les données anonymes et les données personnelles. Ainsi toute trace numérique est potentiellement une donnée personnelle au sens de la loi. Ce lien avec l'origine personnelle des traces peut d'ailleurs apparaître lors de la structuration et de la fouille des traces numériques collectées une fois celles-ci structurées de façon exploitable et ce même lorsqu'une anonymisation des résultats publiés demeure possible.

Le cadre légal de la protection des données à caractère personnel ne s'applique pas uniquement à la publication des résultats. Il fixe des exigences en matière de sécurité informatique, en matière de délais de conservation, et des normes encadrant l'information des enquêtés ou encore le recueil de leur consentement sauf motif d'intérêt légitime dans le cadre d'une mission de service public qui doit être justifiée. Cela peut soulever des problèmes méthodologiques de nature à affecter la qualité de la recherche.

Malgré l'absence d'acteurs issus des SHS dans les discussions sur le cadre dérogatoire applicable à la recherche scientifique prévu par le RGPD, il demeure une possibilité de déroger à certaines de ces règles pour des impératifs matériels ou méthodologiques. Mais la possibilité de recourir à ces règles est soumise à une interprétation dont la compétence, au sens juridique, n'appartient plus exclusivement au chercheur.

Références bibliographiques

Albrecht, Jan Philipp (2015), Hands off our data! Bruxelles : Jan Philipp Albrecht.

Atten, Michel (2013), « Ce que les bases de données font à la vie privée, What databases do to private life », Réseaux, n° 178-179, p. 21-53.

Austin, John Langshaw (1970), Quand dire, c'est faire, Paris : Éditions du Seuil.

Badouard, Romain ; Mabi, Clément ; Monnoyer-Smith, Laurence (2016), « Le débat et ses arènes », Questions de communication, n° 30, p. 7-23.

Bennett, Colin J., ; Raab Charles D. (2003), The Governance of Privacy. Policy Instruments in Global Perspective. Aldershot : Ashgate.

Bergeron, Henri ; Surel, Yves, et Valluy, Jérôme (1998), « L'Advocacy Coalition Framework. Une contribution au renouvellement des études de politiques publiques ? », Politix 11 (41), p. 195-223.

Bigot, Jean-Édouard (2018), Instruments, pratiques et enjeux d'une recherche numériquement équipée en sciences humaines et sociales, Thèse en Sciences de l'information et de la communication, Université de Technologie de Compiègne.

Boullier Dominique (2015A), « Les sciences sociales face aux traces du Big Data », Revue française de science politique, 5-6, p. 805-828.

Boullier, Dominique (2015B), « Pour des sciences sociales de troisième génération (SS3G) : des traces numériques aux répliques », in Menger, Pierre-Michel (coord.), Big Data, entreprises et sciences sociales, Paris : OpenEdition.

Bourdeloie, Hélène (2014), « Ce que le numérique fait aux sciences humaines et sociales. épistémologie, méthodes et outils en questions », *tic&société*, vol. 7, n° 2, p. 7-38

Carré, Dominique et Panico, Robert (2011), Le contrôle social à l'heure des technologies de mobilité et de connectivité, du fichage ciblé des individus au traçage continu des agissements, *Terminal*, n° 108-109, p. 17-31

Collomb, Cléo (2016), Un concept technologique de trace numérique, Thèse en philosophie, Université de Technologie de Compiègne.

Dubois, Vincent (2014), « L'Etat, L'action Publique et La Sociologie Des Champs », *Swiss Political Science Review*, 20 (1), p. 25-30.

Etzioni, Amitai (1999), *The Limits of Privacy*. New York: Basic Books.

Flon, Émilie ; Davallon, Jean ; Tardy, Cécile et Jeanneret, Yves (2009), « Traces d'écriture, traces de pratiques, traces d'identité », *Rétrospective et Perspective, Actes du colloque H2PTM*, Paris : Hermès-Lavoisier.

Foucault, Michel (1975), *Surveiller et punir. Naissance de la prison*, Paris : Gallimard.

Gellert, Raphael (2016), « We Have Always Managed Risks in Data Protection Law: Understanding the Similarities and Differences between the Rights-Based and the Risk-Based Approaches to Data Protection », *European Data Protection Law Review*, vol. 2, n°4, p. 481-492.

Gonzalez Fuster, Gloria (2014), *The Emergence of Personal Data Protection as a Fundamental Right of the EU*, Dordrecht : Springer.

Groupe de travail de l'Article 29 (2007), Avis 04/2007 sur le concept de donnée à caractère personnel, adopté le 20 juin, Bruxelles : Commission européenne

Groupe de travail de l'Article 29 (2014), Avis 05/2014 sur les techniques d'anonymisation, adopté le 10 avril, Bruxelles : Commission européenne

Hondius, Frits Willem (1975), *Emerging data protection in Europe*. Amsterdam : Elsevier.

Jeanneret, Yves (2011), « Complexité de la notion de trace. De la traque au tracé » (p. 59-86), in Galinon-Mélénec, Béatrice (coord.), *L'homme trace. Perspectives anthropologiques des traces contemporaines*, Paris : CNRS Éditions.

Jeanneret, Yves (2013), « Les chimères cartographiques sur l'internet, panoplie représentationnelle de la traçabilité sociale » (p. 235-267), in Galinon-Mélénec, Béatrice ; Zlitni, Sami (dirs.), *Traces numériques : de la production à l'interprétation*, Paris : CNRS Éditions.

Krieg-Planque, Alice (2010), « La formule "développement durable" : un opérateur de neutralisation de la conflictualité », *Langage et société*, n° 134, p. 5-29.

Kuhn, Thomas (1962), *La structure des révolutions scientifiques*, Paris : Flammarion.

Lascoumes, Pierre (2004), « La Gouvernementalité : de la critique de l'État aux technologies du pouvoir », *Le Portique*, n° 13-14, p. 1-15.

Lyon, David (2015), *Surveillance After Snowden*, Cambridge, Mass. : Polity Press.

Mascetti, Sergio ; Monreale, Anna ; Ricci, Annarita ; Gerino, Andrea (2013), « Anonimity : A Comparison Between the Legal and Computer Science Perspectives » (p. 85-115), in Gutwirth, Serge ; Leenes, Ronald ; de Hert, Paul ; Poullet, Yves ; Flinn, Rachel ; Wright, David ; Friedewald, Michael (coord.), European Data Protection: Coming of Age, Dordrecht : Springer.

Merzeau, Louise (2009), « Du signe à la trace : l'information sur mesure », Hermès, n° 53, p. 23-29.

Merzeau, Louise (2013), « L'intelligence des traces », Intellectica, n° 59, p. 115-135.

Mill, John Stuart (1989), On liberty ; with The subjection of women ; and chapters on socialism, Cambridge et New York : Cambridge University Press.

de Montjoie, Yves-Alexandre ; Radaelli, Laura ; Singh, Vivek Kumar ; Pentland, Alex "Sandy" (2015), « Unique in the shopping mall: On the reidentifiability of credit card metadata », Science, vol. 347, n° 6221, pp. 536 - 539

Müller, Pierre (2011), Les politiques publiques, Paris : Presses universitaires de France.

Newman, Abraham (2008), Protectors of Privacy. Regulating Personal Data in the Global Economy, Ithaca : Cornell University Press.

Ohm, Paul (2010), « Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization », UCLA Law Review, n° 57, p. 1701-1777.

Reinach, Adolphe (2004), Les fondements a priori du Droit Civil, Paris : Librairie Philosophique Vrin.

Rieder, Bernhard (2010), « Pratiques informationnelles et analyse des traces numériques : de la représentation à l'intervention », Études de communication, n° 35, p. 91-104.

Rieder, Bernhard ; Röhle, Theo (2012), « Digital Methods: Five Challenges » (p. 67-85), in Berry, David M. (coord.), Understanding Digital Humanities, Londres : Palgrave Macmillan.

Rogers, Richard (2013), Digital Methods, Cambridge : MIT Press.

Rogers, Richard, (2009), The End of the Virtual. Digital Methods, Amsterdam : Vossiuspers UvA.

Rossi, Julien (2015), « Les quatre modèles de Data Protection Officers en Europe » (p. 27-42), in Rasle, Burno (coord.) Correspondant informatique et libertés : bien plus qu'un métier, Paris : AFCDP.

Rossi, Julien (2015), « Données de recherche et vie privée : l'anonymat règle-t-il le problème ? » (p. 319-332), in Rasle, Burno (coord.) Correspondant informatique et libertés : bien plus qu'un métier, Paris : AFCDP.

Rouvroy, Antoinette ; Berns, Thomas (2013), « Gouvernementalité algorithmique et perspectives d'émancipation: Le disparate comme condition d'individuation par la relation ? », Réseaux n° 177, p. 163-196.

Sabatier, Paul A. (1998), « The Advocacy Coalition Framework: revisions and relevance for Europe », Journal of European Public Policy, n° 5:1, p. 98-130.

Simon, Bart (2002), « The Return of Panopticism: Supervision, Subjection and the New Surveillance », Surveillance & Society 3 (1), p. 1-20.

Smadja, David (2012), « La boîte noire de la controverse », Raisons politiques, n° 47 (octobre): 5-11.

Sweeney, Latanya (2000), « Uniquement of Simple Demographics in the U.S. Population », Laboratory for International Data Privacy, Working Paper LIDAP-WP4.

Venturini, Tommaso (2012), « Great Expectations : Méthodes Quali-quantitative et Analyse des Réseaux Sociaux » (p. 39-51), in Fourmentraux, Jean-Paul (dir.), L'Ère Post-Media. Humanités digitales et Cultures numériques, Paris : Hermann.

Venturini, Tommaso ; Latour, Bruno (2009), « Le tissu social/the social fabric : traces numériques et méthodes quali-quantitatives », Proceedings of Future En Seine.

Venturini, Tommaso ; Cardon, Dominique ; Cointet, Jean-Philippe (2014), « Méthodes digitales. Approches quali/quanti des données numériques », Réseaux, n° 188, p. 9-21.

Vitalis, André (2009), « "Informatique et libertés" : une histoire de trente ans », Hermès, n° 53, p. 137-43

Documents téléchargés sur le site web d'ETALAB

Commission nationale de l'informatique et des libertés (CNIL) (2017), « Correspondants Informatique et Libertés (CIL) », Etalab [en ligne], Consulté le 26 juin 2017, http://www.data.gouv.fr/fr/datasets/correspondants-informatique-et-libertes-cil/

Documents issus des deux journées d'étude étudiées

COSTECH (2017), « Synthèse la journée d'étude : "données personnelles en milieu universitaire : quelles questions se poser" ? », Université de technologie de Compiègne [en ligne], Consulté le 20 février 2018, http://www.costech.utc.fr/spip.php?article121

Institut des sciences sociales du politique (ISP) (2017), « Retour en vidéos de la journée d'études "Données personnelles et sensibles, quels enjeux pour la recherche en SHS ?" », Université Paris Nanterre [en ligne] consulté le 20 février 2018, https://www.parisnanterre.fr/actualite-de-la-recherche/retour-en-videos-de-la-journee-d-etudes-donnees-personnelles-et-sensibles-quels-enjeux-pour-la-recherche-en-shs--809751.kjsp

Pasquier, Florent ; Rossi, Julien (2017), Journée d'étude « Données personnelles en milieu universitaire : quelles questions se poser ? », Synthèse de la journée [En ligne], Compiègne - Paris : COSTECH - ISCC, Disponible sur : < http://www.costech.utc.fr/IMG/pdf/synthese_donnees_personnelles_en_milieu_universitaire.pdf >

Ressources ayant contribué à la constitution du corpus documentaire de documents issus du processus d'adoption du RGPD

Albrecht Jan-Philipp (2013), Draft report on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Commission des libertés civiles, de la justice et des affaires intérieures, Parlement européen [en ligne], Consulté le 20 février 2018, http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+COMPARL+PE-501.927+04+DOC+PDF+V0//EN&language=EN

Commission européenne (2009) « Consultation on the legal framework for the fundamental right to protection of personal data », DG HOME, [en ligne], Consulté le 20 février 2018, https://ec.europa.eu/home-affairs/what-is-new/public-consultation/2009/consulting_0003_en

European Digital Rights (EDRi) (2012), EDRi's suggested amendments to the Commission's Proposal for a Regulation on the Protection of individuals with regard to the processing of personal data, and on the free movement of such data (General Data Protection Regulation), Document "EDRi amendments.pdf", Github [en ligne], Consulté le 20 février 2018, https://github.com/lobbyplag/lobbyplag-data/blob/master/raw/lobby-documents/EDRi%20Amendments.pdf

Lobbyplag (n.d.), « lobbyplag-data/raw/lobby-documents at master », Github [en ligne], Consulté le 20 février 2018, https://github.com/lobbyplag/lobbyplag-data/tree/master/raw/lobby-documents

Parlement européen (n.d.), « Fiche de procédure : 2012/011(COD) », Observatoire Législatif [en ligne], Consulté le 20 février 2018, http://www.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2012/0011(COD)&l=fr

Corpus de jurisprudences citées

Verdict du Tribunal constitutionnel fédéral allemand (BVerfG) du 15 décembre 1983 (BVerfG Urteil vom 15. Dezember 1983 Az. 1 BvR 209/83, 1 BvR 484/83, 1 BvR 420/83, 1 BvR 362/83, 1 BvR 269/83, 1 BvR 440/83) (Décision « *Volkszählungsurteil* »)

Cour de justice de l'Union européenne (CJUE) 16 juillet 2015 «ClientEarth contre EFSA» Aff. C-615/13P

Cour de justice de l'Union européenne (CJUE) 20 décembre 2017 «Peter Nowak contre Data Protection Commissioner» Aff. C-434/16

Entretiens et enregistrements cités dans l'article

Genvresse, Sophie, dans : Institut des sciences sociales du politique (ISP) (2017b), « Débat 1 : "Données personnelles et sensibles, quels enjeux pour la recherche en SHS ?" », Youtube [en ligne], Consulté le 20 février 2018, https://www.youtube.com/watch?v=NyrbzTesLTg&t=132s

Georges, Fanny (2018), Entretien réalisé à Paris le 29 janvier 2018

Skowronska, Kaja (2017), Entre responsabilité éthique et contrainte formelle – le traitement des données confiées aux chercheurs lors des enquêtes qualitatives. Notes de lectures non publiées. Intervention à la journée d'étude du 12 janvier 2017 « données personnelles en milieu universitaire : quelles questions se poser ? ». Enregistrement audio disponible en ligne sur le site web du COSTECH, et consulté le 20 février 2018, http://www.costech.utc.fr/IMG/mp3/08_kaja_skowronska_entre_responsabilite_ethique_et_contrainte_formelle.mp3