Analyser les données du web politique : un défi pour les sciences sociales

Samy Cohen, Nonna Mayer, Julien Boyadjian, Stéphanie Wojcik, Camille Escudé-Joffres

SÉMINAIRE

SciencesPo
CENTRE DE RECHERCHES INTERNATIONALES

SciencesPo
CENTRE D'ÉTUDES EUROPÉENNES ET DE POLITIQUE COMPARÉE

Les sciences sociales en question :
grandes controverses épistémologiques et méthodologiques

Compte rendu de la 42e séance

Analyser les données du web politique : un défi pour les sciences sociales

12 avril 2018

La 42e séance du séminaire[1] est consacrée aux problèmes méthodologiques que pose l'analyse des données qui circulent sur la toile. Les millions de messages politiques (*tweets, hashtags, blogs*) qui s'y échangent tous les jours mettent les sciences sociales au défi. Comment sélectionner au sein de cette masse de données volatiles, nombreuses, souvent anonymes ? Comment analyser ces dernières ? Dans quelle mesure sont-elles représentatives et extrapolables au reste de la population ? Les deux intervenants, spécialistes du web politique, tentent de répondre à ces questions.

Nonna Mayer commence par les présenter. Julien Boyadjian, maître de conférences à Sciences Po Lille/CERAPS, a fait sa thèse sur les messages politiques circulant sur Twitter, publiée sous le titre *Analyser les opinions politiques sur Internet. Enjeux théoriques et défis méthodologiques* (Paris, Dalloz, 2016). Il a publié récemment plusieurs articles consacrés à la méthode, notamment avec Aurélie Olivesi et Julien Velcin *Le web politique au prisme de la science des données. Des croisements disciplinaires aux renouvellements épistémologiques*[2].

[1] Compte-rendu réalisé par Camille Escudé (CERI).

[2] *Réseaux*, 4 (204), 2017, pp. 9-31. Voir aussi : « Les conditions de scientificité des *big data* en science politique, *Revue française de science politique*, 67(5), 2017, pp. 919 - 929, « L'analyse quantitative des médias sociaux, une alternative aux enquêtes déclaratives ? *Questions de communication*, 31(1), 2017, pp. 111 - 135 et « Comment étudier les réseaux sociaux ? », in Xavier Marc et Jean-François Tchernia (dir.), *Etudier l'opinion* [2e édition augmentée], Grenoble, PUG, 2018 (à paraître).

La discutante est Stephanie Wojcik. Maîtresse de conférences à l'Université Paris Est Créteil/CEDITEC, responsable du réseau de recherche Démocratie électronique (DEL), elle s'intéresse à la participation politique et à la citoyenneté en ligne. Elle a notamment publié La citoyenneté numérique. Perspectives de recherche (avec Fabienne Greffet)[3] et « Analyser la participation politique en ligne : des traces numériques aux pratiques sociales » (avec Gersende Blanchard et Simon Gadras)[4].

Julien Boyadjian

Julien Boyadjian centre son intervention sur le travail effectué dans le cadre de sa thèse de doctorat. Son projet est né alors qu'il effectuait un stage dans un cabinet d'études qui réalise des sondages d'opinion classiques. Il s'intéresse alors à la façon dont on peut renouveler les enquêtes d'opinion dans le domaine commercial grâce aux nouveaux outils numériques. A l'époque, il était surpris par le peu de réflexion existant en science politique sur l'usage possible des nouvelles données et des nouveaux outils, utiles pourtant pour travailler sur les opinons politiques des citoyens. Sa question de départ est donc celle des promesses et des limites des données récoltées sur le web par rapport à celles obtenues lors d'enquêtes traditionnelles en sciences sociales.

Publier sur les réseaux sociaux est aujourd'hui un phénomène social massif. La moitié de la population française est inscrite sur les réseaux sociaux et un tiers publie des contenus. Ces contenus numériques peuvent être appréhendées comme des indices sociologiques, qui offrent des informations sur les représentations et les attitudes des internautes, et donc comme un matériau d'enquête.
A partir des années 2000-2010 se développent des nouveaux outils pour utiliser de manière scientifique les données du web qui permettent de collecter ces données de manière automatisée (messages, nombre de *followers*, de *retweets*, etc.). On peut également stocker et archiver ces données. Enfin, on peut les analyser à partir d'outils volumétriques, lexicométriques, sémantiques. Le chercheur dispose d'une part d'une une masse formidable de données et de l'autre d'outils volumétriques et sémantiques qui permettent de les collecter et de les interpréter.
En science politique et dans la recherche sur les opinions publiques, ces données numériques présentent quatre grands avantages par rapport aux données d'enquêtes déclaratives traditionnelles.

Tout d'abord, elles ne sont pas générées en situation de laboratoire. Le chercheur est donc certain de ne pas imposer sa problématique. De plus, les enquêtes par questionnaire ont du mal à atteindre les opinions stigmatisées ou extrêmes. Ici, on a accès aux réactions des individus sur le moment, sans qu'ils aient à rechercher dans leur mémoire. Pour autant, les données ne sont pas complètement spontanées, elles restent dans le cadre d'interactions sociales. Les auteurs de ces données savent que leurs écrits vont être lus et cela peut impacter la teneur des messages. De plus, ces données peuvent être formatées en termes de nombre de caractères sur Twitter, Facebook défavorise l'anonymat, etc.
En raison de leur abondance, ces données peuvent être mobilisées pour des analyses tant

[3] *Réseaux*, 2 (184-185), 2014, pp.125-159
[4] In Christine Barats (dir.), *Manuel d'analyse du web en sciences humaines et sociales,* Paris, Armand Colin, 2013, pp.166-180.

qualitatives que quantitatives. Elles permettent de cartographier des communautés d'intérêt et d'opinion et de dégager des influences réciproques à partir des *like, follow, retweet* des auteurs des données. Cela permet par exemple de faire apparaître l'homogamie politique : on se cite entre groupes semblables. Ces données sont traçables et permettent d'étudier les opinions de manière dynamique et dans le temps.

Julien Boyadjian a été séduit par ces promesses mais s'est rapidement trouvé confronté à la limite des données. Celles-ci ne sont en effet pas forcément associées à des auteurs clairement identifiés et on ne connaît pas forcément leurs propriétés sociodémographiques. Pour cette raison, on parle souvent des « internautes » et des « réseaux sociaux » de manière globale, mots qui ne veulent pas dire grand-chose. Julien Boyadjian a donc eu la volonté d'aller plus loin pour savoir ce qu'écrivent les internautes selon leur position sociale, leur genre etc. Certains auteurs disent qu'il faut faire le deuil de ces données sociodémographiques et qu'il faut chercher au-delà. Pour Dominique Boullier[5], on peut faire des sciences sociales « de troisième génération » dans lesquelles les phénomènes de propagation compteraient plus que les caractéristiques des individus. Julien Boyadjian refuse cette position et affirme que ces données numériques peuvent dire beaucoup si on parvient à les rattacher au monde réel.

Julien Boyadjian s'est donc trouvé face à un défi méthodologique : comment tirer profit des apports heuristiques des *big data*, tout en parvenant à contourner leur principale limite ? Autrement dit, comment parvenir à analyser des données non générées par un dispositif d'enquête contrôlé, tout en situant socialement et politiquement leurs auteurs ?
Le dispositif qu'il a mis au point consiste à constituer un panel représentatif d'usagers d'un réseau social dont on aura identifié les propriétés sociales et politiques à l'aide d'un questionnaire comportant une vingtaine de questions d'ordre sociodémographique et dont on analysera l'activité de publication à l'aide d'un logiciel pendant un an. Cela permet d'analyser des *verbatims* non suscités par le chercheur mais socialement situés. Julien Boyadjian précise que le choix de Twitter est un choix par défaut. Il voulait au départ s'intéresser à Facebook dont l'audience est plus importante : 60% des internautes sont sur Facebook, 11% seulement sur Twitter à l'époque (en 2011). Facebook permet donc l'accès à un public plus diversifié en termes d'âge et de classes sociales et constitue un matériau plus riche. Le chercheur s'est cependant heurté à des difficultés quand il s'est agi d'accéder aux comptes. Les membres de Twitter sont moins diversifiés mais la très grande majorité de leurs comptes sont accessibles.

L'enquête a comporté plusieurs phases :

1. Constitution à partir d'un logiciel de collecte de *tweets* d'une base de données exhaustive de 248 928 comptes Twitter ayant publié au moins un *tweet* politique entre le 1er et le 31 mars 2012.
2. Extraction d'un échantillon de 20 000 comptes soumis à une analyse d'éligibilité (individu unique, français, compte ouvert).
3. Administration d'un questionnaire aux 10 299 comptes éligibles.
4. Archivage de la totalité des *tweets* publiés par les 608 répondants (6%).
5. Constitution d'un panel de contrôle de 620 « non-répondants ».

[5] *Sciences humaines et sociales, troisième génération* https://shs3g.hypotheses.org/ .

Julien Boyadjian a essayé de voir dans quelle mesure ces répondants étaient représentatifs des non répondants. Il a tenté d'obtenir le maximum d'informations *via* le contenu des *tweets* et leur description. Il a ainsi pu identifier le sexe, l'âge, le lieu de résidence, le niveau de diplôme (*via* LinkedIn), la catégorie socioprofessionnelle, l'appartenance politique des personnes. Il a ensuite mis en parallèle les résultats du panel répondant et du panel non répondant. Une fois que les individus ont donné leur accord, il a aspiré les *tweets* pendant une période de onze mois, du 1er mars 2013 au 31 janvier 2014. Il regrette que la conjoncture n'ait pas été particulièrement favorable, c'est-à-dire qu'elle n'ait pas été une période électorale intense. Néanmoins, le mariage pour tous, l'affaire Cahuzac, la liaison entre François Hollande et Julie Gayet ont été au cœur de l'actualité.

Au total, le panel des répondants a fourni 840 251 *tweets* publiés sur la période, dont 81 606 contenant des occurrences politiques, soit 9,7% du total. Le panel des non répondants a fourni 1 120 896 *tweets*, dont 42 940 *tweets* politiques, soit 3,8% des *tweets* publiés. Les non-répondants ont donc produits plus de *tweets* mais apparaissent moins politisés.
La première série de résultats porte sur la sociologie des auteurs de *tweets* politiques. Comparée à l'ensemble des Français la population qui s'exprime sur Twitter est plus politisée et participe davantage, elle milite, elle vote, etc. La seule différence notable est qu'elle est très jeune : les 18-24 ans y sont surreprésentés à l'inverse de ce que l'on observe à l'échelle de l'ensemble des Français où les plus âgés sont surreprésentés dans les activités partisanes et politiques. Certains pensent que les réseaux sociaux peuvent ré-enchanter la participation politique des individus mais les jeunes qui s'expriment politiquement sur Twitter ont le même profil que ceux qui votent aux élections. Les jeunes sont surreprésentés sur Twitter mais ne sont pas pour autant représentatifs de la jeunesse française. Ils sont issus de la jeunesse étudiante des cycles longs (universités, classes préparatoires et grandes écoles). Environ 30% de ces jeunes sortent de classe préparatoire et de grandes écoles.

Une autre série de résultats provient du croisement des données d'activité avec des données sociodémographiques. Julien Boyadjian a identifié quatre variables qui influent sur la fréquence de messages politiques :

- L'âge : on s'exprime plus quand on est âgé (malgré la surreprésentation des jeunes) ;
- L'intérêt pour la politique ;
- Le militantisme ;
- Le degré d'activité sur le réseau : plus on est actif, plus on parle de politique.

Enfin, l'analyse dans le temps de la production des messages a permis d'identifier que la production de *tweets* est fortement liée à la production médiatique générale. Plus les médias parlent de politique, plus on *tweete*. Des pics de production correspondent à des événements politiques fortement médiatisés. Toutes les catégories d'usagers parlent davantage de politique en période de forte intensité médiatique.

Julien Boyadjian identifie pour finir plusieurs limites à ce dispositif. Celui-ci est extrêmement coûteux en temps et peut difficilement être reproduit en dehors d'un travail de doctorat. Il ne permet pas d'étudier des corpus exhaustifs, à la différence du *big data* dont l'exhaustivité constitue l'un des principaux avantages . Il est encore nécessaire d'échantillonner les données. Une question éthique se pose enfin : peut-on analyser ces données sans le consentement des internautes ? Dans le cadre de ses recherches, Julien Boyadjian a contacté la Commission nationale informatique et libertés (CNIL), qui lui a demandé d'obtenir l'accord explicite des enquêtés pour le panel des répondants. Les données Twitter ne sont

pas considérées comme « publiques » mais seulement comme « accessibles » et la CNIL part du principe qu'il n'est pas certain que les utilisateurs sachent faire la différence. La Commission a donc demandé au chercheur d'introduire volontairement dans ses données un pourcentage de 5% d'erreur afin de brouiller les pistes. Julien Boyadjian précise cependant que la CNIL reste bienveillante par rapport aux chercheurs et surveille surtout ce que les entreprises privées font des données du web.

Julien Boyadjian conclut en proposant des pistes alternatives pour travailler sur les données du web « non situées ». Si peu d'internautes dévoilent leur profession ou leur niveau de diplôme, beaucoup révèlent leurs goûts en matière de culture, de consommation, d'actualité, de politique, etc. (fonctions « j'aime », « *follow* », « partage », etc.). Ces goûts nous apprennent beaucoup sur la position sociale et le positionnement politique des internautes. Il est donc possible de constituer des catégories d'internautes selon leurs intérêts et leurs goûts. Cette piste est déjà investie par Facebook, qui a déposé un algorithme pour identifier de manière automatique les internautes via les contenus qu'ils déclarent aimer ou qu'ils partagent, etc.

Stephanie Wojcik

Stephanie Wojcik (Université Paris Est Créteil/CEDITEC) se réjouit que la science politique se saisisse à la fois des problématiques et des questions de méthodologie liées aux phénomènes du web politique. Le numérique infuse dans toutes les dimensions de l'activité sociale et inaugure un nouvel âge de la traçabilité.
Se pose tout d'abord la question de la pluridisciplinarité. Les méthodes qu'on emploie en tant que chercheur peuvent-elles être appliquées à l'analyse du web politique ? A-t-on besoin de méthodes nouvelles ou bien peut-on importer les méthodes classiques des sciences humaines et sociales, comme la traditionnelle enquête par questionnaire ? Les masses de travaux existant sur Twitter, qui concernent les contenus discursifs produits, se heurtent régulièrement aux problèmes de la linguistique automatisée ou du traitement automatique des langues (qui associe la linguistique et l'informatique et qui consiste à développer des logiciels ou des programmes informatiques pour traiter de manière automatisée des données linguistiques). Ainsi, les logiciels ne prennent en pas en considération toutes les caractéristiques sémiotiques. Une étude du Pew Research Center dont un compte rendu a été publié sur *Wired* montre que la plupart des liens qui renvoient vers des sites populaires sur Twitter sont émis par des *bots*. Les chercheurs du Pew ont automatisé la collecte des liens sur Twitter à l'aide d'un outil appelé Botometer – un *machine learning* – qui « prédit » et catégorise tel ou tel compte comme « automatique »[6]. En utilisant ses critères (fréquence des *tweets*, etc.), Botometer estime que le compte de Barack Obama est un compte automatique.

Evidemment, certaines nuances échappent à l'analyse robot : le sarcasme, l'ironie, pourtant fréquents dans le langage politique. Ce type d'analyse se heurte donc à des difficultés classiques. A cet égard, le travail sur les idéologies de Julien Longhi, linguiste à Cergy Pontoise, réalisé à partir d'une analyse des *tweets* émis pendant les élections municipales de 2014 constitue un bon exemple [7]. Julien Longhi pensait que l'idéologie pouvait se repérer

[6]"Most links to popular sites on Twitter come from bots", *Wired*, 9 April 2018, https://www.wired.com/story/twitter-bots-links/
[7] https://www.u-cergy.fr/fr/_plugins/mypage/mypage/content/jlonghi.html.

dans des formes énonciatives (absence de connecteur argumentatif, absence de nom propre, conjugaison au présent, etc.) dont l'analyse peut être automatisée mais à l'épreuve des faits, il s'est aperçu que cela ne fonctionnait pas. Autrement dit, certains objets, qu'ils soient ou non discursifs, peuvent être rétifs aux traitements automatisés ou bien nécessiter de très lourds traitements.

A propos de la collecte des contenus issus de différentes plateformes (Facebook, Instagram, etc.), il n'est désormais plus possible de procéder avec des copier-coller d'écrans. Par exemple, le chercheur est obligé de passer par des outils qui permettent de recueillir un certain pourcentage de *tweets*, selon des critères exclusivement définis par Twitter. Cette difficulté qui n'est cependant pas propre au web est redoublée par le caractère volatile et mouvant des contenus du web. Autre exemple, il est très difficile de récupérer les sites Internet des partis politiques dans leur intégralité. La Bibliothèque nationale de France les archive mais ne conserve pas tout ce qui est accessible par au moins trois clics de l'internaute sur les sites, ce qui limite le travail du chercheur.

Stephanie Wojcik revient ensuite sur la distinction opérée entre données issues du web et celles recueillies par exemple à travers des questionnaires. En réalité, il n'existe pas de données « naturelles »[8] qui contrasteraient avec les données issues de protocoles d'enquête qui seraient donc artificiellement produites. Les architectures digitales tendent à laisser leur marque sur les données qu'elles permettent de collecter. De fait, plusieurs études montrent clairement à quel point les argumentations développées par les internautes diffèrent sensiblement selon les différentes plateformes sur lesquelles ils expriment leur opinion.

Outre les questions d'outils, on rencontre des problèmes liés à la définition de la problématique. Est-il toujours pertinent de recueillir des grandes masses de données ? Rappelons simplement que la constitution du corpus dépend de la problématique, et que la taille du corpus n'est pas synonyme de qualité ou d'exhaustivité et ne constitue pas une garantie de pertinence et d'objectivité. Se pose alors la question de l'extrapolation de ce qui est observable en ligne à ce qui est observable hors ligne. Les résultats présentés confirment ce qu'on sait du web politique : on y trouve plutôt des gens intéressés et politisés qui tirent profit des nouvelles ressources pour s'informer et se mobiliser. Globalement, on observe un élargissement de la fracture existant entre les plus politisés et les moins politisés.

Il est également difficile pour un chercheur en sciences humaines et sociales de maîtriser la question du codage pour récupérer du contenu sur Facebook. Il faut pour cela réaliser un petit programme informatique et les chercheurs en sciences humaines et sociales ne sont pas nécessairement en mesure de connaître le codage et encore plus le *machine learning*. De plus, les compétences en codage sont le plus souvent l'apanage des hommes que des femmes. De nouvelles hiérarchies se créent entre les chercheurs qui possèdent des compétences en informatique et les autres qui sont dans la nécessité de les développer au fur et à mesure de leur recherche.

Le dernier problème évoqué est celui de l'éthique. Les politiques en vigueur sur ce sujet diffèrent selon les pays. Stephanie Wojcik explique avoir eu affaire à l'équivalent canadien de notre Commission nationale informatique et libertés, celle-ci exerce un contrôle beaucoup plus strict que nous le faisons en France. Les contenus auto-publiés (*tweets, posts* de

[8] Lewis K., "Three fallacies of digital footprints", *Big Data and Society*, 2(2), 2015.

Facebook, billets de blogs, …) émanent-ils d'une activité privée ou publique ? La notion d'*accountability*[9] est parfois avancée par certains chercheurs : il faut être en mesure d'apporter des réponses claires si l'on vous demande ce que vous avez fait des données récoltées.

Stephanie Wojcik termine en posant quelques questions à Julien Boyadjian.

- Outre les renseignements socio-professionnels sur les auteurs de *tweets,* une analyse des messages postés a-t-elle été effectuée ?
- Peut-on dire que tout *tweet* correspond à une opinion ?
- Comment peut-on mesurer la participation politique ?

Julien Boyadjian

Julien Boyadjian remercie pour les questions qui lui ont été posées. Il précise que dans son travail de thèse, il n'a pas travaillé à partir des outils lexicométriques. D'une part, parce qu'il ne les maîtrise pas et d'autre part parce que les outils à disposition étaient peu performants. Il a mis en place une typologie par rapport à une question de départ et a codé les messages. Il a ensuite tiré au sort des messages, pour savoir quel était l'objet du *tweet* : un objet politique, une prise de position politique ou bien le rapport à l'éthique, à l'homme politique, son physique, sa compétence, etc. Ensuite, il a effectué un codage manuel à partir d'une grille d'analyse thématique.

Pour le chercheur, il ne s'agit pas de savoir si le web renforce la participation politique des citoyens ; il se demande plutôt si Twitter nous apprend autre chose des opinions politiques que ce que permettent les sondages traditionnels. Par exemple, on reproche souvent aux enquêtes de répondre aux préoccupations des commanditaires. Il est donc intéressant d'observer si les utilisateurs des réseaux sociaux ont des préoccupations différentes de celles que reflètent les questions posées par les sondeurs. Sur Twitter, les usagers parlent des mêmes choses que les journalistes dans les médias. Julien Boyadjian a pu constater que les mêmes thématiques sont abordées dans son panel et dans celui de l'ensemble des hommes et des journalistes politiques présents sur Twitter. Twitter n'est donc pas représentatif de la population française dans son ensemble mais des journalistes et des étudiants des grandes écoles. Donc, il a raisonné à l'envers : il est parti des individus pour remonter à leurs comptes Facebook et Twitter. En conclusion, le web 2.0 ne favorise pas la participation de ceux qui se sentent exclus de la sphère politique. Cependant, Julien Boyadjian précise que sur Facebook, on n'étudie pas tout à fait les mêmes choses. Sur Twitter, les sources et les sujets sont plutôt sérieux ; sur Facebook, ils sont plus populaires. Les deux applications ne visent pas les mêmes publics et les contenus qu'elles partagent diffèrent.

Quant à savoir si le *tweet* vaut opinion, l'*a priori* de départ était de considérer Twitter comme un observatoire de l'opinion politique. Or les *tweets* sont aussi souvent bien autre chose, des billets d'humeurs, des partages d'informations, l'expression de goûts et de dégouts, des formes de mises à distance de la politique, etc.

[9] Dumez Hervé, « De l'obligation de rendre des comptes ou *accountability* », *Annales des Mines-Gérer et comprendre*, 2008/1 (n° 91), pp. 4-8. https://www.cairn.info/revue-gerer-et-comprendre1-2008-1-page-4.htm.

Discussion avec le public

Josiane Joüet (Paris 2 CARISM) félicite Julien Boyadjian pour la méthodologie très intelligente mise en œuvre et représentative des utilisateurs de Twitter. Elle souligne que les utilisateurs qui publient (des contenus en ligne) sont les plus politisés. Les utilisateurs qui ne produisent pas eux-mêmes mais lisent ou partagent des contenus sont nombreux et plus difficile à saisir. Elle apprécie l'effet d'agenda du travail de Julien Boyadjian. Les personnes s'emparent des thèmes politiques médiatisés, qu'ils aillent ou non dans le sens de leurs propres opinions, pour provoquer des réactions durant un ou deux jour(s). Josiane Joüet interroge Julien Boyadjian, a t-il observé de tels phénomènes ?
Julien Boyadjian est d'accord sur la distinction entre utilisateur politisé et personne politisée tout court. Il explique travailler sur un nouveau panel d'étudiants de classe préparatoire qui ne twittent pas forcément sur l'actualité mais aussi sur leur quotidien. La question de savoir ce qui favorise l'expression d'opinions politiques n'est pas encore résolue. Pour ce faire, on peut faire des entretiens classiques semi-directifs pour comprendre ce qui fait qu'on s'exprime sur les réseaux. La politisation n'est pas une condition suffisante, il n'y a pas forcément un usage politisé du réseau social. Pour répondre à la deuxième question, il explique avoir choisi d'appréhender les *tweets* de manière isolée plutôt que les *tweets* analysés dans les discussions ou les réponses dans un fil de discussion.

Joyce Bessis (CERI) explique avoir commencé le même travail, mais à l'envers. Elle travaille sur les acteurs politiques français et sur la façon dont ils utilisent les réseaux sociaux. Elle est étonnée d'entendre parler d'éthique pour la première fois et elle demande quelle est la problématique exacte de la thèse. Florence Nocca (CEE), qui rédige également une thèse mobilisant les données Twitter, revient sur les questions éthiques, soulignant que la DSI de Sciences Po indique qu'il lui faut demander le consentement individuel de chaque député. Elle précise que les mentions légales de Twitter soulignent que tout usager accepte que ses publications soient utilisées à des fins de recherche.
Benjamin Ooghe-Tabanou (Médialab) affirme être également très étonné par la réponse de la CNIL. Il n'a jamais entendu parler des « 5% d'erreurs » à ajouter. Il précise également que le terme de *big data* est souvent utilisé à mauvais escient et qu'il concerne en vérité des données beaucoup plus importantes par leur taille. Il pose la question des outils et des logiciels utilisés.

Julien Boyadjian précise qu'il est allé rendre visite à la CNIL dans le cadre de sa thèse qui faisait partie d'un projet ANR en 2013-2014, à une époque où sa thèse était presque terminée et où la CNIL elle-même était en pleine réflexion sur ces questions. Il n'y avait pas encore de jurisprudence et la seule recommandation qui lui a été faite a été d'inclure un pourcentage d'erreurs.
A propos des compétences en informatique, Julien Boyadjian précise que dans le cadre du projet ANR Magiweb, il a travaillé avec des informaticiens. Pour la première fois, deux milieux se sont rencontrés et les interactions n'ont pas toujours été simples, ce qui nous amène à la question suivante : travaille-t-on en véritable collaboration avec des informaticiens ou bien le personnel technique doit-il répondre à la demande qui lui est faite ? Il faudrait plus d'ingénieurs de recherche qui maîtrisent les codes informatiques dans les laboratoires.

La problématique de départ était de savoir si le web et ses outils et données permettaient de contourner les principales limites des sondages d'opinion, en cernant mieux les préoccupations des enquêtés. La conclusion est que Twitter n'est pas le meilleur terrain pour

ce faire parce que ses utilisateurs sont très politisés et proches par leurs opinions des milieux politiques et journalistiques. Il est donc difficile d'extrapoler à partir d'eux et d'appliquer les résultats obtenus aux internautes en général.

Claire Andrieu (Centre d'histoire de Sciences Po) demande quelle définition de la politique a été utilisée et quels mots-clés « politiques » ont été privilégiés par Julien Boyadjian. Celui-ci lui répond qu'il a utilisé les noms des candidats aux élections et une liste de cent mots clés qui se réfèrent à la vie politique institutionnelle (Assemblée nationale, nom des partis, noms des personnalités politiques, etc.). Il a utilisé d'autres mots pour tester (chômage, grève, etc.) et il est retombé sur les mêmes personnes. Pour cette population spécifique des *tweeters* politiques il ne pense pas être passé à côté de beaucoup d'individus.

Nonna Mayer (CEE) évoque à propos de la distinction public-privé un problème du même ordre rencontré lors de la rédaction d'un ouvrage sur le Front national qu'elle a codirigé avec Sylvain Crépon et Alexandre Dézé, *Les faux semblants du Front national*[10]. Dans son chapitre sur les usages frontistes du net, Julien Boyadjian voulait intégrer deux captures d'écran prises sur des sites frontistes, Les Patriotes et le Rassemblement bleu Marine. L'éditeur, en l'occurrence les Presses de Sciences Po, a subordonné leur publication à une autorisation officielle du Front national. Après bien des péripéties, Nonna Mayer raconte comment elle a réussi à joindre Florian Philippot, alors vice-président, au téléphone et à obtenir son accord. Benjamin Ooghe-Tabanou ajoute que dans le cas de l'utilisation de Google, les données de Google Trend sont accessibles au grand public. Il précise qu'il a apprécié la méthodologie de Julien Boyadjian, qui se distingue du recours massif et sans nuances au quantitatif.

Nonna Mayer rebondit sur ce qui vient d'être dit pour souligner l'importance du travail collaboratif entre chercheur.e.s en sciences humaines et sociales et chercheur.e.s en sciences de l'information. Benjamin Ooghe-Tabanou propose à tous de participer à l'atelier qui a lieu à 14h30 tous les deuxièmes mardis du mois, le Metlax qui permet des collaborations de ce type. Les inscriptions sont ouvertes sur le site du Medialab[11].
Stephanie Wojcik conclut la séance en rappelant que si les problèmes d'éthique sont très importants, les chercheurs collectent pour l'heure les données sans se poser trop de questions. A cet égard, les politiques sont très différentes selon les universités et les pays. En France où il n'existe pas encore de directive précise, chacun bricole dans son coin. Bref, les questions d'éthique sont toujours sans réponse mais invitent les chercheurs à la réflexivité.

[10] Julien Boyadjian, « Les usages frontistes du web », in Sylvain Crépon, Alexandre Dezé, Nonna Mayer (dir.), *Les faux-semblants du Front national. Sociologie d'un parti politique*, 2015, Paris, Presses de Sciences Po, pp. 141-159.
[11] https://medialab.sciencespo.fr/fr/atelier/ .