Politique et Sociétés

Politique et Sociétés

Conjuguer principes et pratiques éthiques au temps des données socionumériques : le nécessaire dialogue

Tania Gosselin

Volume 41, numéro 3, 2022

URI : https://id.erudit.org/iderudit/1092346ar
DOI : https://doi.org/10.7202/1092346ar

Aller au sommaire du numéro

Éditeur(s)

Société québécoise de science politique

ISSN

1203-9438 (imprimé)
1703-8480 (numérique)

Découvrir la revue

Citer ce document

Gosselin, T. (2022). Conjuguer principes et pratiques éthiques au temps des données socionumériques : le nécessaire dialogue. *Politique et Sociétés*, *41*(3), 241–246. https://doi.org/10.7202/1092346ar

Conjuguer principes et pratiques éthiques au temps des données socionumériques : le nécessaire dialogue

Entretien avec Yannick Dufresne, chercheur à l'Université Laval qui s'intéresse aux données socionumériques depuis une décennie. Les propos ont été recueillis par Tania Gosselin (Université du Québec à Montréal). Les échanges se sont déroulés en septembre et octobre 2021.

Q : Comment peut-on définir le « big data » ?

YD : Avant de lancer la Chaire [de leadership en enseignement des sciences sociales numériques] il y a quatre ans, on m'avait suggéré d'utiliser « données massives » dans son nom ; je préfère parler de données numériques, car le problème n'est pas toujours le volume, c'est-à-dire le côté « *big* » des données. La gestion de ce type de données pose des défis techniques, mais il est important de considérer aussi les défis théoriques et éthiques qui découlent de l'exploitation des nouvelles données numériques en sciences sociales. Je suis heureux de voir que, dans le monde vaste et confus où certains parlent de données numériques et d'autres de données massives, une certaine cohésion commence à prendre forme autour de la notion de sciences sociales numériques. Dans un récent numéro de *Nature*[1], David Lazer et ses collègues parlent eux aussi de trois défis – conceptuel, méthodologique et éthique – en lien avec les sciences sociales numériques.

La définition la plus classique est celle des « 3V » : le premier « V » est pour Volume. C'est l'aspect « *big* » – qui est relatif à la discipline, car ce qui est considéré comme étant un gros volume n'est pas la même chose en science politique et en génétique. Avant, c'était ce qui attirait toute l'attention. Et ça d'ailleurs un peu contribué à développer un certain fétichisme de la méthode. Beaucoup de travail a été fait depuis pour gérer le volume. Le

1. Lazer, David, Eszter Hargittai, Deen Freelon *et al.* 2021. « Meaningful Measures of Human Society in the Twenty-first Century. » *Nature* 595 : 189-196. https://doi.org/10.1038/s41586-021-03660-7. Consulté en juillet 2021.

deuxième « V » est celui renvoyant à la Variété des structures et de la nature de ce type de données ; je lui accorde beaucoup d'importance, car la variété des données – de sondage, de géolocalisation, textuelles, relationnelles, physiologiques – est un défi majeur. Nous ne sommes plus dans le paradigme qu'on pourrait décrire comme « rectangulaire », avec des bases de données « lignes–colonnes ». Sur Twitter, par exemple, il y a aussi des liens entre les gens ; il s'agit de mettre les données en forme pour aller chercher l'information et s'assurer qu'elles puissent contribuer à la connaissance. Le troisième « V » est pour Vélocité, le débit constant de données.

Certains parlent d'un quatrième « V » pour Véracité ou Validité. La validité des données qu'on collecte et analyse est d'une importance primordiale ; sinon, on passerait non seulement à côté du grand potentiel de ces données, on risquerait même de nuire en fondant des décisions sur des conclusions erronées. Toutefois, et c'est un point de vue personnel, la validité des conclusions découle directement de la logique de la recherche scientifique qui guide la manière de relever les défis théoriques ou conceptuels. Il faut que ces données puissent avoir une validité qui permette de comprendre quelque chose qu'on ne saisissait pas avec d'autres types de données.

Q. Qu'est-ce que cela change dans la manière de faire des sciences sociales ?

YD : Mon champ d'expertise est l'opinion publique. Je vois que les journalistes sont sur Twitter, que les parlements rendent accessibles en ligne les discours [des élus] et de plus en plus d'informations ; il y a les initiatives de gouvernements ouverts dans les villes et ailleurs. Depuis toujours, on a besoin de connaître l'opinion des citoyens, mais les espaces citoyens numériques se fragmentent ; les gens répondent de moins en moins aux sondages. Comment trouver le moyen de continuer à capter les opinions citoyennes ? Je ne suis peut-être pas objectif à ce sujet puisque c'est ce que j'étudie, mais je ne vois pas d'avenir dans l'étude de l'opinion publique sans plonger dans les possibilités qu'offre le numérique. Les traces numériques laissées par les citoyens ne fournissent pas seulement des données d'une richesse inédite pour mieux décrire et comprendre leurs attitudes et comportements ; la nature même de l'opinion publique a été modifiée par les outils de l'ère numérique. On risque fort de ne pas bien cerner cette opinion si on se limite aux données et méthodes traditionnelles. Cela dit, nous allons toujours avoir besoin de l'esprit humain pour guider et encadrer nos recherches sur le plan théorique. L'interprétation demeure aussi, pour l'instant, une chasse gardée de l'intelligence humaine. Un algorithme qui permet d'analyser un texte ne remplacera peut-être jamais un humain ; mais un humain ne peut pas lire un milliard de tweets. Si des humains pouvaient lire les milliards de Tweets avec une procédure de codage traditionnel, ce serait l'idéal ! Quand on se rabat sur l'analyse automatique, on perd nécessairement quelque chose.

Mais l'exhaustivité et la reproductibilité que permettent ces méthodes ne sont pas négligeables. Les méthodes d'analyse textuelle automatique rendent carrément faisable l'étude de certains phénomènes. Afin de bénéficier du plein potentiel des nouvelles données numériques en sciences sociales, il est primordial de réfléchir aux limites et aux apports de chacune des technologies et des expertises concernées de manière interdisciplinaire. La science est à son meilleur quand elle est une entreprise sociale.

Q : Qu'en est-il du défi éthique à l'ère numérique ?

YD : Quoique fondamental, le défi éthique n'est pas intrinsèque à la définition des données numériques. C'est nous [en recherche] qui décidons de donner une importance à cette dimension pour des raisons qui dépassent la simple accumulation de connaissances. Par exemple, la recherche a une tradition de consentement des participants, mais quand on va chercher des données sur Twitter [sur des comptes publics et pour lesquelles le consentement n'est pas requis] et que ça peut permettre de faire des prédictions du vote des personnes uniquement en considérant les retweets [sans même avoir besoin de ce que les utilisateurs expriment sous forme de contenu], c'est certain que ça soulève des nouvelles questions qu'il faut considérer selon une perspective d'éthique de la recherche. Il est donc primordial d'amorcer ces réflexions avec les comités d'éthique.

Avant, une sorte de quasi-monopole des outils était détenu par les universités et quelques grosses compagnies et il était facile de vérifier [ce qui se faisait]. Maintenant, mon étudiant peut faire de la reconnaissance faciale dans son sous-sol ! On sait aussi qu'il y a des pouvoirs étrangers qui peuvent avoir des intentions qui n'ont rien à voir avec la recherche. Quand je vois des entreprises ou même des particuliers utiliser ces données à différentes fins sans formation scientifique ou cadre éthique, il est certain que je souhaite que les universités jouent un rôle actif dans la recherche de solutions pour mieux protéger les citoyens. C'est un immense défi. Je n'ai pas de réponse [à toutes les questions]. Sans aucun doute, l'application de contraintes éthiques devrait idéalement être élargie au-delà du domaine de la recherche. Je suis un peu inquiet du fossé pouvant se creuser entre les capacités d'analyses d'intérêts privés et celles des chercheurs qui veulent mieux comprendre l'influence et les effets potentiellement néfastes de ces nouvelles données sur la société. Les défis éthiques posés par la collecte, le stockage et l'utilisation de ces nouvelles données en sciences sociales provoquent une réflexion qui doit absolument impliquer un large spectre d'expertises techniques, théoriques et éthiques. Il faudra trouver un moyen de protéger les citoyens tout en ne minant pas l'important potentiel de ces données pour la recherche et ses bénéfices pour la société. Les universités doivent évidemment jouer un rôle central dans cette réflexion dont les répercussions vont bien au-delà du monde de la recherche.

Q : Dans un récent article publié dans Nature*, des chercheurs qui travaillent avec des données socionumériques mentionnent des questions liées au défi éthique, dont la ré-identification par recoupage des données et le « spillover » – le fait que des informations sur un internaute puissent fournir des informations au sujet d'autres internautes en raison de la nature des données en réseaux. Comment fait-on pour protéger la confidentialité des données dans ce contexte ?*

YD : Le défi de la ré-identification est de taille. Il faut s'assurer que même des bases de données qui ne contiennent que peu d'information identificatoire directe ne puissent pas être croisées de manière à ré-identifier les sujets. En effet, si plusieurs bases de données, collectées indépendamment, contiennent des informations communes pour l'analyse (sociodémographique ou autre), il peut devenir assez facile de faire le lien entre ces bases pour compléter ou détailler les informations des individus qui auraient répondu à ces différentes études. Même si les bases de données de départ, prises indépendamment, n'auraient pas permis l'identification des individus, une base de données ainsi fusionnée pourrait contenir les informations nécessaires à la ré-identification de ces individus.

Q : Aurait-on besoin de retravailler l'Énoncé de politique des trois conseils : Éthique de la recherche avec des êtres humains – EPTC 2 (2018) pour l'adapter aux nouveaux types de données ? Il y a aussi la Politique de gestion des données de recherche[2] *qui s'ajoute à ce cadre. Ou faudrait-il plutôt mettre l'accent sur la mise en œuvre du cadre par les comités d'éthique de la recherche ?*

YD : Je n'ai pas de réponse simple à cette question. Du côté des comités, il va sans dire que l'évaluation des recherches en numérique complexifie leur tâche en posant de nouveaux défis. Il faudra anticiper un moment d'adaptation, mais je suis confiant que leur expertise essentielle saura nous guider vers des solutions raisonnables. Mon espoir repose aussi sur le développement et la diffusion d'une sophistication numérique et logicielle dans l'ensemble des organisations et des institutions dédiées à l'évaluation de projets numériques. Je pense en particulier aux projets d'infrastructures de recherche – comme un logiciel dédié à la collecte, au stockage et au partage de données sécurisées – et aux outils de transfert des connaissances. Dans ce contexte, on ne peut plus penser la recherche comme un *one shot deal* : on va chercher un montant, on collecte des données, on analyse, puis on ferme les livres et on recommence. Construire un logiciel représente des centaines de milliers de dollars […] Les chercheurs ont besoin de soutien, et pas seulement financier. Je suis loin d'être un spécialiste de l'éthique. Or, à mon avis, l'idéal

2. https://www.science.gc.ca/eic/site/063.nsf/fra/h_97610.html. Consulté en juin 2021.

n'est pas seulement de conventionner les principes de l'éthique en recherche, mais aussi de construire de manière à ce qu'un autre chercheur avec un projet de recherche similaire au mien ait accès à une infrastructure commune permettant la sécurisation des données. Il faut éviter de reconstruire toute la machinerie de l'anonymisation des données pour 50 000 ou 75 000 dollars pour chaque projet. Il faudrait s'entendre pour déterminer ce qu'il nous faut et construire la « tuyauterie » logicielle en collaboration avec les comités éthiques, les chercheurs, des gens en génie logiciel ; il faut sécuriser le tout pour l'ensemble de la communauté universitaire. Tout le monde dit qu'il faut arrêter de travailler en silo, qu'il faut davantage d'interdisciplinarité – c'est facile à dire mais c'est difficile à faire. Travailler en silo dans le développement logiciel et de structures, c'est complètement sous-optimal et inefficace ! Cet exercice, en plus d'être extrêmement fastidieux, entraîne un énorme gaspillage de ressources, de temps et d'argent lors du développement de différentes solutions aux mêmes problèmes.

On ne peut pas demander à des chercheurs en science politique ou en communication d'être spécialistes de ce type de structures – on n'a pas ces compétences-là ! J'ai été chanceux d'acquérir une expertise particulière presque par hasard en répondant aux défis posés par mes projets de transfert de connaissances. J'aime l'idée de construire des infrastructures de recherche pouvant servir à d'autres chercheurs par la suite. J'utilise souvent la métaphore de la construction de maisons : bien que les projets de maison puissent prendre différentes formes et couleurs, il serait absurde de repenser un système d'égouts pour chacune de ces maisons. De la même façon, les projets de recherche peuvent avoir différentes formes et couleurs, mais les besoins liés au stockage, à la sécurité et au partage des données sont souvent communs. Lorsque plusieurs chercheurs partagent une même infrastructure, des économies d'échelle sont évidemment possibles, en temps et en argent. De plus, l'intérêt partagé favorise le maintien de ces infrastructures, qui est aussi un enjeu majeur. J'imagine aussi que, du même coup, ce type d'infrastructure commune faciliterait l'évaluation éthique des projets de recherche axés sur le numérique. Pour poursuivre avec la métaphore de la maison, il faudrait que ceux qui font des plans d'architecte travaillent avec ceux qui font des solages, et que les comités d'éthique supervisent le tout en plus de l'aménagement urbain et du zonage !

Si l'on vise une recherche transdisciplinaire, on aura besoin non seulement de normes mais aussi d'outils communs, des outils logiciels qu'on pourra évaluer et parfaire, qui seront plus pérennes et allégeront le fardeau pour les prochains chercheurs qui veulent développer des projets numériques […] Il faut un investissement et de la reconnaissance pour le solage et les égouts. Pour ma part, il est certain qu'en tant que politologue, je préférerais pouvoir consacrer davantage de temps à l'étude de mes objets

de recherche, au défi conceptuel plutôt qu'au défi technique. J'ai bon espoir qu'on soit sur la bonne voie.

Q : À quel niveau serait-il souhaitable de développer cette « tuyauterie » ?

YD : Selon moi, il faut surtout être cohérent. C'est la même idée que pour l'Énoncé ; ça prend des principes de base communs qui s'appliquent à tous les chercheurs […] Des employés de grandes compagnies de technologie ont parfois leur nom sur des articles scientifiques. Ces données-là sont-elles des données secondaires ? Sont-elles soumises à un examen par un comité d'éthique ? Mon inquiétude est que les contraintes liées à la collecte de données dans le milieu scientifique créent des incitatifs qui nuisent ultimement aux objectifs de protection des citoyens pour lesquels nous avons créé ces contraintes. Comme le mentionnent Lazer et ses collègues, quand les chercheurs s'intéressent à des thèmes qui intéressent moins les propriétaires des données, ou qu'ils ne veulent pas collaborer avec ces propriétaires pour éviter de faire des compromis, cela pourrait par exemple les inciter à ne pas respecter les conditions d'utilisation d'une plateforme. Il faut un contrepoids aux données régies par la propriété privée et à la recherche de type un peu « Far West » qui peut se faire dans le privé. C'est une des conséquences possibles du manque d'uniformisation des contraintes éthiques dans les différents secteurs utilisant ces nouvelles données sociales […] Ce problème ne se limite pas seulement à la comparaison entre le privé et l'université, mais peut aussi être soulevé quant à la comparaison entre les universités. Si le potentiel des données numériques n'est pas accessible de manière uniforme dans différentes universités, il va y avoir des disparités immenses. Dans un monde idéal, tous les chercheurs et utilisateurs de données seraient soumis aux mêmes règles ! Les cadres éthiques sont évidemment là pour protéger les citoyens car il y a eu des dérives importantes ; il ne faut jamais, jamais retourner là. Mais l'objectif de protéger tous les citoyens ne me semble pas atteint si les règles ne s'appliquent qu'au monde universitaire ou s'il existe des disparités à ce sujet entre les universités.

La recherche universitaire est soumise à de hauts standards éthiques, c'est nécessairement un des éléments qui la définit. Ça doit être davantage reconnu par tous, selon moi. J'ai quand même un petit malaise quand je vois de la recherche faite par des compagnies privées partagée dans les médias avec le plus souvent une méthodologie incomplète. Je suis peut-être idéaliste, mais j'aimerais que tous les citoyens puissent interpréter la présence de logos d'universités sur des projets de transfert de connaissance et des résultats de recherche comme un gage de rigueur scientifique et éthique. C'est peut-être une bonne façon de sensibiliser les citoyens à la valeur de leurs données personnelles tout en favorisant la collecte de données éthique, ce qui nécessairement valorisera encore davantage la recherche universitaire.