SCIENCES XXL

Ce que l'abondance et la diversité des données numériques font aux sciences sociales

Journées d'étude

Jeudi 16 et vendredi 17 mars 2017

Ined - SAGE

APPEL A COMMUNICATION

Données de l'Internet, données administratives, données de capteurs, fichiers de gestion (du personnel, d'adhérents, d'anciens élèves), etc. ... le volume d'informations numériques disponibles a fortement cru ces dernières années. Une fois passée la fascination pour ces gisements de données en apparence gratuites, abondantes et immédiatement disponibles, les questions soulevées par l'usage scientifique de ces masses numériques sont nombreuses. Certaines sont classiques, d'autres moins, d'autres enfin se posent avec une acuité renouvelée.

Le plus souvent produites à des fins gestionnaires, administratives ou commerciales, les informations collectées ne sont pas toujours utiles pour les chercheur-e-s, en tous cas pas immédiatement ; quoique massives, elles ne sont pas toujours plus précises, ni plus fiables que des données d'enquête classique ; les questions d'accès et de propriété des données compliquent le travail ou modifient la place des chercheur-e-s, qu'elles-ils soient demandeur-e-s ou, de manière croissante, sollicité-e-s par les producteurs pour les traiter.

L'objectif de ces journées d'étude est de prolonger ces réflexions théoriques, méthodologiques, épistémologiques et éthiques pour savoir ce que l'abondance et la diversité des données numériques font aux pratiques de recherche en sciences sociales, en France comme à l'étranger. Les communications pourront s'appuyer sur des récits d'expériences ou des restitutions de recherches déjà menées, en vue d'expliciter ces pratiques de manière réflexive, ou de saisir d'éventuels changements de ces dernières au cours du temps. Des présentations sur la production ou les usages que les administrations ou les entreprises font des données numériques pourront aussi être proposées. Les comparaisons (historiques et/ou avec d'autres disciplines scientifiques) sont particulièrement bienvenues afin de mieux cerner l'éventuelle nouveauté de la situation actuelle.

Les propositions de communication, **à renvoyer avant le 22 novembre 2016,** pourront s'inscrire dans l'un des 3 axes suivants :

1. *Données cherchent chercheur-e-s : conditions d'accès et d'usage des données*

La multiplication de sources de données, par des « nouveaux » producteurs le plus souvent situés hors du monde académique, modifie parfois substantiellement

la position des chercheur-e-s. Pour accéder aux informations, ces dernier-e-s doivent engager des négociations avec les propriétaires des données, parfois sans connaître en amont leur qualité, leur nature et la quantité / représentativité des informations disponibles. Cela passe souvent par la signature d'un contrat de recherche ou d'une convention de partenariat dans lesquels les contraintes imposées peuvent être importantes.

La présence accrue du droit dans la relation d'enquête n'est pas qu'une affaire de juriste. De manière croissante, les chercheur-e-s sont aussi invité-e-s, en échange de l'accès aux données, à les traiter pour le compte de l'organisation, à adopter des modes de restitution séparés (académiques et experts), voire à mettre en place un partenariat où les résultats seront valorisés par l'organisme producteur. Les chercheur-e-s peuvent aussi être directement démarché-e-s par des institutions publiques ou privées pour travailler sur leurs données, moins sur la base de leur connaissance approfondie d'un champ scientifique que de leur compétence technique en matière de traitement de données. Ces nouvelles conditions modifient-elles les pratiques de recherche (par exemple, les manières de concevoir son objet de recherche, les modes de reconnaissance entre pairs, la manière de constituer les équipes de recherche et de répondre à des appels d'offre, *etc.*) et dans quel sens ? Comment la collaboration entre ces producteurs de données et le monde de la recherche s'organise-t-elle ?

Une fois obtenues, l'usage et la conservation des données sont également encadrés par un ensemble de règles, certaines légales, d'autres contractuelles, d'autres encore auto-imposées. Beaucoup d'informations utilisées dans les sciences sociales tombent sous le coup de la loi Informatique et Libertés de 1978 sur les données personnelles, et doivent donc être collectées avec parcimonie, voire pas du tout. Certaines doivent être détruites après coup, ce qui bouscule a priori le principe de cumulativité de la science et les possibilités de « revisite » des terrains et objets d'enquête. Dans d'autres cas, ce sont les institutions propres au champ scientifique (les comités d'éthique ou les *review boards*) ou les normes propres au milieu qui viennent encadrer ces usages. Comment les chercheur-e-s négocient-elles/ils ces contraintes institutionnelles ? Quelles transformations cela implique t-il dans la pratique concrète de la recherche, la division du travail scientifique (de la négociation à la restitution) ?

Ces données posent aussi la question du coût d'accès pour les chercheur-e-s. En apparence gratuites, ces dernières peuvent nécessiter un accès sécurisé et/ou payant, des infrastructures de stockage et de sauvegarde, des logiciels puissants pour les traiter, voire l'embauche de personnel pour les préparer et les mettre en forme, augmentant les inégalités d'accès entre institutions et laboratoires. Enfin, dans le cadre d'accords signés avec les producteurs, les chercheur-e-s peuvent être contraint-e-s quant aux modalités de restitution des résultats, leurs publications académiques peuvent par exemple être encadrées par des clauses de notification, un droit à la primo-publication, etc. Parallèlement, les données qualitatives collectées dans le cadre d'enquête multi-méthodes posent la question de leur statut, des conditions de leur anonymisation comme de leur propriété. Comment ces demandes modifient-elles le travail scientifique, depuis sa conception jusqu'à sa restitution ?

2. Des objets aux savoirs constitués : objets, disciplines et théories

Parmi la masse de données numériques disponibles, beaucoup proviennent de capteurs individuels (disposés par exemple sur les téléphones portables, les ordinateurs, les montres, etc.) ou de fichiers propres à des organisations et n'ont pas été conçues pour les besoins de la recherche. La critique habituelle selon laquelle ces données ne pourraient pas servir au travail scientifique doit être nuancée : les chercheur-e-s travaillent depuis longtemps sur des données qui n'ont pas été conçues par et pour elles-eux (archives administratives, bases de données d'entreprises, etc.). Mais cette critique doit aussi être précisée. Que peut-on faire avec ces données : à quelles questions permettent-elles de répondre auxquelles on ne pouvait pas répondre avant ? Et à quelles conditions (redressement, contrôle) ? Dans quelle mesure sont-elles compatibles ou complémentaires avec d'autres sources et méthodes d'enquête ? Les travaux articulant (ou comparant) différentes sources de données, d'enquête (produites par des chercheur-e-s ou la statistique publique) et de gestion (produites par des administrations, des entreprises, des associations, etc.) sont de ce point de vue bienvenus ; ils s'efforceront de discuter l'apport empirique de ces bases numériques et leurs effets éventuels sur la chaîne de production de la recherche (lien avec les services d'appui à la recherche et les ingénieurs de recherche, etc.).

La question des objets de la connaissance se pose à nouveau avec acuité. L'abondance de certaines sources tend à privilégier certains sujets ou questions au détriment d'autres, pour lesquels les données sont moins directement disponibles - qu'ils soient moins facilement quantifiables, statistiquement marginaux, économiquement non rentables. Les objets de recherche ont-ils déjà évolué, et si oui dans quel sens ? Les catégories utilisées pour décrire le monde social changent t-elles avec la modification des sources ? Les pratiques des disciplines comme les frontières classiques entre elles en sont-elles affectées ? Les communications portant sur les effets de la multiplication des données sur les savoirs produits sont bienvenues. Outre les questions de méthodes ou les formes de connaissance favorisées, les présentations pourront réfléchir aux stratégies mises en place par les chercheur-e-s pour gérer l'abondance ou en tirer profit.

Enfin, les résultats doivent être considérés. Le phénomène des « big data » a été initialement célébré pour la capacité que cette masse numérique aurait à faire entrer les sciences sociales dans une nouvelle ère théorique et/ou prédictive. Mais quelles en sont les réalisations concrètes ? Et au prix de quels déplacements ? L'abondance transforme t-elle vraiment les manières de faire et si oui, dans quel sens ? S'oriente-t-on par exemple vers une forme d'hyperempirisme au détriment d'autres formes de conceptualisation ? Et plus généralement, les savoirs produits augurent-ils d'une nouvelle hiérarchie des manières de faire et des objets en sciences sociales ?

3. Les outils de la connaissance : collecter, compter, nettoyer, analyser

Les communications pourront enfin mettre en avant les outils mobilisés pour collecter, stocker, explorer et traiter ces données. Cette question en apparence technique et logistique est lourde d'enjeux scientifiques et épistémologiques. Quels sont les savoirs produits, et en quoi diffèrent-ils de ceux issus des modes de collecte classiquement utilisés en sciences sociales (questionnaires, archives, etc.) ? Les recherches empiriques menées sur des lieux de production de ces données (de la conception des capteurs téléphoniques par exemple aux plateformes d'agrégation de données), éventuellement armées par les outils de l'anthropologie des sciences, sont bienvenues, tout comme le sont les communications qui s'intéressent à la production des données ouvertes (*open data*) et aux expériences de collecte collective (*crowdsourcing*).

Les outils disponibles pour traiter ces données évoluent aussi. La masse de données modifie par exemple les usages possibles de la statistique et leur articulation avec les autres techniques d'enquêtes ethnographiques. L'abondance (ou la complétude) de certaines bases rend ainsi potentiellement caduque toute une série de mesures classiquement utilisées, comme les tests d'inférence. Les données elles-mêmes sont souvent présentées dans des formats qui rendent plus difficile le recours aux techniques classiques (que ce soit parce qu'elles présentent un nombre important de valeurs manquantes, un déséquilibre entre variables et individus aboutissant à la production de tableaux difficilement exploitables, *etc.*). Quelles stratégies adopter pour faire face à ces données, et pour quels résultats ? Le récent retour en grâce de techniques comme l'apprentissage statistique [*machine learning*] montre bien certains déplacements en cours. Nées à la croisée des mathématiques et de l'informatique, et désormais largement utilisées dans différents espaces sociaux pour explorer et prédire les comportements des consommateurs, ces méthodes bouleversent-elles les pratiques de recherche et d'analyse ?

On note aussi une demande croissance de compétences destinées à collecter, apparier et nettoyer ces données (redresser des valeurs manquantes, les rendre cohérentes, *etc.*). Cette valorisation récente de techniques issues de l'informatique renforce t-elle l'opposition (trop) conventionnelle entre qualitativistes et quantitativistes, ou déplace t-elle ces lignes en ouvrant une troisième ligne de front méthodologique (quali, quanti, « ordi ») ?

➔ L'objectif des journées d'étude est de réunir les chercheur-e-s de différentes disciplines (démographie, sociologie, économie, histoire, science politique, géographie, gestion, anthropologie, archéologie *etc.*), mobilisant des données abondantes. Les communications auront pour objet les reconfigurations de la recherche suite à cette massification et à cette diversification, qu'il s'agisse de la transformation des méthodes ou des objets, des savoirs produits, des relations avec les enquêté-e-s et la place des chercheur-e-s, ou encore des enjeux juridiques et éthiques.

Afin de créer une discussion collective et transversale sur ces sujets, deux formats de présentation sont proposés :

- **Des présentations courtes** (moins de 10 minutes). Basées sur un retour d'expérience, elles traiteront un des points de l'appel à communication.
- **Des présentations plus longues** (20 minutes environ). Sur la base d'une enquête approfondie ou d'une comparaison entre sources de données, elles proposeront une réflexion approfondie sur l'une ou plusieurs questions de la journée.

Les journées alterneront les présentations courtes et longues, et seront complétées par une table ronde et une conférence plénière.

Les propositions de communication doivent être présentées selon le format suivant (2 pages environ) :
Titre de la communication, nom(s) du ou des auteur(es), adresse(s) postale(s) et électronique(s) et appartenance(s) institutionnelle(s). La proposition de communication précisera le type de communication ainsi que la base empirique et/ou les méthodes employées.

Calendrier

- Clôture de l'appel à communication le **22 novembre 2016**
- Décision du comité scientifique et notification : **18 décembre 2016**
- Journées : **Jeudi 16 et vendredi 17 mars 2017**

Adresse de contact : scienceXXL@gmx.com

Lieu : Ined, salle Sauvy, 133 boulevard Davout, 75020 Paris

Comité d'organisation
Anne Lambert (Ined-CMH), Étienne Ollion (Université de Strasbourg-SAGE), Arnaud Bringé (Ined – chef du service des méthodes statistiques).

Comité scientifique
Andrew Abbott (sociologue - Université de Chicago), Didier Breton (démographe - Université de Strasbourg-SAGE), Dominique Cardon (sociologue - Sciences Po), Béatrice Cherrier (économiste - Université de Caen), Arthur Charpentier (économiste - Université de Rennes 1), Emmanuel Didier (sociologue - CNRS), Frédéric Lebaron (sociologue - ENS Cachan), Eva Lelièvre (démographe - INED), Claire Lemercier (historienne - CNRS), Alain Trognon (statisticien - INSEE), Laurent Toulemon (démographe - INED), Florence Weber (anthropologue - ENS Ulm).