Les enjeux éthiques du « tournant numérique » des sciences sociales. S'approprier les outils pour la production et la gestion des documents d'enquête

**Thibaut Rioufreyt (Docteur en science politique,
Chercheur associé au laboratoire Triangle)**

Conférence dans le cadre de la 3ème journée de réflexion « Enjeux éthiques autour de la production et la gestion des données en sciences sociales », journée organisée par la Commission d'éthique de la recherche de la Faculté des sciences de la société de l'Université de Genève

26 février 2019

UniMail, Université de Genève (UNIGE)

INTRODUCTION (1)

Les effets de l'essor des technologies numériques sur les SHS

- Nouveaux « terrains » : réseaux sociaux (Twitter, Facebook, LinkedIn, etc.), forums et groupes de discussion web, Fab-Lab, hackerspaces, etc.
- Nouveaux objets d'étude – makers, cyberhacktivistes, participation politique online, etc.
- Nouveaux types de « données » (corpus web) dans des volumes inédits (Big Data)
- Nouveaux outils et/ou méthodes d'analyse : exploration/extraction de données (data mining), analyse de réseaux, simulation, etc.

Cette conjonction de nouveaux « terrains », de nouveaux objets, de nouveaux types et volumes de données et de nouvelles techniques et méthodes est à l'origine de l'émergence d'un nouveau champ d'étude au sein des sciences sociales : les Internet Studies (Dutton 2013).

INTRODUCTION (2)

Certains auteurs diagnostiquent ou appellent ainsi de leurs vœux l'émergence de sciences sociales numériques sous différentes appellations et labels : « e-social science » (Halfpenny Peter et Procter Rob 2010), « digital social research » , « computational social science » (Lazer et al. 2009) ; (Cioffi-Revilla 2010), « digital social science » (Spiro 2014), « digital cultural heritage » (Cultural Heritage Informatics Initiative 2009), etc.

D'autres annoncent la naissance d'une nouvelle épistémologie pour les sciences sociales. Le numérique ne serait ainsi pas simplement à l'origine de nouveaux champs d'étude (comme les Internet Studies) au sein des sciences sociales, ou de nouvelles subdivisions (entre les sciences sociales numériques et les autres), mais à l'origine d'une nouvelle manière de faire science sociale appelée à remplacer les SHS telles qu'on les a connues jusque-là.

INTRODUCTION (3)

Un autre aspect du tournant numérique des sciences sociales, basique et pourtant fondamental sous-estimé, voire ignoré de la littérature :

Les outils numériques fondamentaux que chacun.e de nous utilise pour la production et la gestion des données : dictaphones numériques ou smartphones utilisés pour enregistrer des entretiens ou des séances d'observation, logiciels de bureautique (traitement de texte, tableur, etc.), bases de données, espaces de stockage en ligne comme Dropbox ou Google Drive où l'on met ses documents, messageries comme Gmail ou Outlook par lesquelles on envoie des messages à nos enquêtés, aux membres de l'équipe de recherche, repositories où l'on dépose nos enquêtes en vue de leur partage et/ou de leur archivage, etc.

INTRODUCTION (4)

La thèse défendue ici est à la fois simple et radicale : la plupart des outils que nous utilisons vont à l'encontre des droits des enquêté.e.s concernant leurs données personnelles.

Le RGPD comme la plupart des textes législatifs ou réglementaires nationaux antérieurs spécifient bien que les données personnelles, plus encore celles à caractère sensible, ne sauraient être communiquées à un tiers sans l'autorisation de la personne, encore moins être manipulées à des fins commerciales. Pourtant ces outils ont été précisément conçus pour collecter le maximum de données.

Autrement dit, on a un problème. Et un gros.

INTRODUCTION (5)

Le propos de cette conférence est double.

1) il s'agira de montrer en quoi la plupart des outils les plus utilisés posent problème au regard de la protection des données personnelles à la fois des chercheur.e.s et des enquêté.e.s et, d'autre part, en quoi les outils *open source* constituent souvent une alternative utile et efficace en la matière.

2) Seront abordés différents outils numériques qui non seulement ne communiquent pas les données à des tiers mais peuvent aider le(s) chercheur.e(s) à protéger celles-ci.

Pour cela, cette conférence sera construite en suivant chaque étape de la vie des documents d'enquête : (co)production, traitement et analyse des matériaux (1), stockage et sauvegarde (2) et partage et archivage (3).

PLAN DE LA CONFÉRENCE

1) PRODUCTION, TRAITEMENT ET ANALYSE DES MATÉRIAUX

1.1) La collecte/production des matériaux

1.2) Le traitement des matériaux

1.3) L’analyse

2) STOCKAGE ET SAUVEGARDE

2.1.) Sauvegardes et récupérations

2.2) Accès et sécurité des données

3) LE PARTAGE ET ARCHIVAGE DES DOCUMENTS D’ENQUÊTE

3.1) Comment partager les documents d’enquête ?

3.2) Favoriser le référencement des enquêtes : les identifiants numériques pérennes

3.3) Mettre en place un dispositif de partage contrôlé des données à usage restreint

1) PRODUCTION, TRAITEMENT ET ANALYSE DES MATÉRIAUX

1.1) La collecte/production des matériaux (1)

Quelques exemples d'outils numériques à cette étape de l'enquête :

1) L'usage des smartphones pour enregistrer des entretiens ou des séances d'observation ethnographique

2) Les logiciels pour la passation de questionnaires en ligne

3) La collecte documentaire sur des sites web d'individus ou de groupes étudiés

4) La collecte de corpus web issus des réseaux socio-numériques (Twitter, Facebook, LinkedIn, etc.) et des forums de discussion online

1.1) La collecte/production des matériaux (2)

Les smartphones sont sans doute le pire outil de traçage et de collecte de données. Il existe bien quelques smartphones *open-source* (comme Librem 5 de l'états-unien Purism, Ubuntu Touch ou le français Eelo). Toutefois, leur développement reste pour l'instant confidentiel et se heurte à de sérieux problèmes, aboutissant parfois à de retentissants échecs (comme Firefox OS). Cela étant dit : il existe une série de pratiques permettant de limiter les frais :

- Désactivation de la géolocalisation
- Installation de moteurs de recherche *open-source* (comme Firefox)
- Contournement des packages d'applications comme GooglePlay ou AppelStore
- Non-connexion aux réseaux sociaux
- Chiffrement des SMS, via Signal
- Chiffrement des emails, via des dispositifs comme Enigmail
- Installation d'un VPN (*Virtual Private Network*) ou réseau privé virtuel

1.1) La collecte/production des matériaux (3)

D’autres chercheur.e.s utilisent des logiciels pour la passation de questionnaires en ligne, à l’instar de SurveyMonkey, SurveyGizmo, GoogleForms. Là encore, il convient de recourir à des solutions techniques garantissant la protection des données. En ce domaine, les choses sont néanmoins un peu moins pires en raison du succès qu’a connu Limesurvey, logiciel opensource, en SHS.

1.1) La collecte/production des matériaux (4)

La collecte documentaire sur des sites web d'individus ou de groupes étudiés devient également une méthode de recueil des données très pratiquée par les chercheur.e.s. Dans ce cadre, une attention toute particulière doit être portée sur deux outils : le navigateur web et le moteur de recherche.

Concernant les premiers, les mauvais élèves en termes de protection des données sont sans surprise ceux proposés par les GAFAM, comme Google Chrome, Safari, Internet Explorer/Edge. À l'inverse, deux d'entre eux, open source, se distinguent particulièrement en la matière : Firefox (moyennant une utilisation effective des fonctionnalités proposées comme la « navigation privée ») et Tor.

1.1) La collecte/production des matériaux (5)

Utiliser des moteurs de recherche respectueux des données personnelles (Startpage qui constitue une couche libre sur Google, Duck Duck Go, Qwant, etc.) à travers une série de fonctionnalités qu'ils proposent :

- Masquage du numéro IP de l'ordinateur, empêchant toute forme de web tracking
- aucun stockage des données : Il n'y a littéralement aucune donnée concernant l'utilisateur sur les serveurs de ces infrastructures.
- Blocage des cookies
- Pas de filtres : Les moteurs de recherche comme Google ou Bing se servent de vos habitudes de recherche pour vous fournir les résultats qu'ils estiment être ceux que vous souhaitez, vous piégeant essentiellement dans une chambre d'écho de résultats.

1.1) La collecte/production des matériaux (6)

Dernière méthode pour laquelle le choix de l'outil s'avère déterminant pour la protection des données : la collecte des corpus web issus des réseaux socio-numériques et des forums de discussion online, etc.

Les chercheur.e.s travaillant sur de telles données doivent être particulièrement attentifs aux techniques de crawling et de scrapping qu'ils vont utiliser.

Les controverses sur les Big Data et les multiples scandales sur l'usage des réseaux lors de campagnes électorales (Cambridge Analytica pour le référendum sur le Brexit et la campagne Trump, l'usage de Whatsapp par l'équipe de Jair Bolsonaro lors des dernières élections présidentielles au Brésil) font que c'est sans doute l'un des domaines les plus sensibles en la matière.

1.2) Le traitement des matériaux (1)

Il existe toute une série d'outils numériques assistant le/la chercheur.e en SHS dans le traitement des matériaux. Certains d'entre eux soulèvent de vrais problèmes en termes de protection des données personnelles, d'autres au contraire peuvent aider à cette dernière.

Parmi les outils problématiques, on peut citer par exemple des outils très spécifiques et un peu périphériques par rapport au cœur des activités de la recherche, comme les services en ligne de réduction de la taille des documents PDF ou les logiciels de conversion de formats de fichiers.

Mais cela concerne également des outils plus centraux dans le travail scientifique, comme par exemple les logiciels de transcription et les outils d'aide à l'anonymisation.

1.2) Le traitement des matériaux (2)

Utilisation croissante de logiciels pour les aider dans la transcription des enregistrements audio ou vidéo des entretiens ou des observations qu'ils ont réalisés (Rioufreyt, 2018). Or, tout comme le choix de l'OS ou de son navigateur web, la localisation du programme n'est pas simplement une question technique ; elle est tout autant politique et déontologique. Les logiciels de transcription en ligne, à l'instar de Transcribe ou oTranscribe, posent en l'occurrence de sérieux problèmes. À titre d'exemple, on sait que Google sur lequel fonctionne Transcribe duplique l'ensemble des données des utilisateurs dans ses data centers. C'est écrit noir sur blanc dans les Conditions générales d'utilisation, que la plupart du temps l'utilisateur se garde de lire. oTranscribe garantit davantage le respect du caractère privé des données : les fichiers audio et les transcriptions ne quittent jamais l'ordinateur. Toutefois, il est lié à une application Google sur smartphone.

1.2) Le traitement des matériaux (3)

Il existe des logiciels d’aide à l’anonymisation qui peuvent s’avérer très utiles pour la protection des données personnelles.

1) Les données de recherche sous forme de texte ou d'images peuvent être constituées de fichiers créés par les enquêté.e.s eux-mêmes. Le risque d'identification à partir des métadonnées est particulièrement élevé dans ces cas. Les métadonnées peuvent être supprimées à l'aide d'éditeurs de texte ou d'images courants (MS Office, Libre Office, Explorateur de fichiers Windows, Photoshop, GIMP, Irfanview, par exemple). Il existe également des programmes spécialement conçus pour supprimer les données EXIF (par exemple, Easy Exif Delete), qui facilitent la suppression des métadonnées masquées. Voir le The Metadata Anonymization Toolkit.

2) Il existe en outre des outils d’aide à l’anonymisation de texte, à l’instar du T ext Anonymization Helper Tool ou de données quantitatives comme Amnesia.

1.3) L’analyse des matériaux (1)

Les logiciels d’analyse ou d’aide à l’analyse doivent également être pensés du point de vue de la protection des données personnelles.

Quelques exemples :

- l’utilisation de logiciels d’analyse synchronisés depuis des serveurs en ligne hébergés et gérés par des sociétés privées
- l’utilisation de programmes de crawling/scrapping pour la collecte des données web
- l’utilisation de cartes Google Maps pour la représentation cartographique des données. Alternative : OpenStreetMap (OSM)
- Etc.

Voir le guide « Software development with Data Protection by Design and by Default » élaboré par la Norwegian Data Protection Authority

2) STOCKAGE ET SAUVEGARDE

2.1) Sauvegardes et récupérations (1)

Comment stocker et sauvegarder les documents d'enquête ? Quelques conseils :

- Sauvegardez les données et les différentes versions régulièrement
- Stockez au moins une copie de sauvegarde dans un autre emplacement physique
- Assurez-vous que les copies de sauvegarde n'ont pas été corrompues, par exemple, en utilisant des sommes de contrôle (checksums)
- Faites du versioning, c'est-à-dire sauvegarder les différentes versions successives d'un même document, en inscrivant la date et le numéro de la version dans l'intitulé du fichier. Cela peut s'avérer utile pour le(s) chercheur(s) en cas de fichier corrompu, d'erreur manuelle ou de travail collectif mais aussi pour les réutilisateurs afin de comprendre le déroulement de l'enquête et son contexte.
- Des copies de travail et des copies de sauvegarde doivent toujours être créées pour tous les fichiers liés aux documents d'enquête.
- Créer un dossier distinct pour chaque jeu de données dans lequel les fichiers de données, les informations de description et tous les autres fichiers liés aux données seront sauvegardés. Il est conseillé de sauvegarder ces fichiers au même emplacement que les métadonnées.
- L'enregistrement et la copie sur différents supports sont généralement faciles si la taille des données ne crée pas de restrictions. Les données d'enquête sont rarement assez volumineuses pour poser problème, mais les données d'enregistrement ainsi que le matériel audio et vidéo peuvent nécessiter des mesures exceptionnelles.
- Recourir aux services de sauvegarde automatique plutôt que de se fier à des procédures manuelles. Attention : la synchronisation n'est pas une sauvegarde ! Si vous supprimez un fichier sur une machine synchronisée avec d'autres, il sera supprimé sur tous !

2.1) Sauvegardes et récupérations (2)

Où stocker les documents d’enquête ? Quelques conseils :

- Éviter les supports optiques (CD, DVD, Blue-ray) et avec une mémoire non volatile (cartes mémoire ou clés USB, par exemple). Les périphériques externes les plus sûrs pour conserver les fichiers de données sont les disques durs externes (HDD).
- Étant donné que les lecteurs de disque dur sont susceptibles de tomber en panne, il est conseillé de copier les mêmes données sur plusieurs disques durs et/ou d'utiliser un support supplémentaire pour la sauvegarde (voir ci-dessous).
- Compléter ces sauvegardes par une solution de stockage fiable, gérée et proposée par les services informatiques de votre Université ou par des infrastructures de recherche dédiées, à l’instar d’Huma-Num en France. Au cours d'un projet de recherche, les membres du projet sont généralement responsables du stockage des documents d’enquête. L'archivage a généralement lieu à la fin du projet. Il est toutefois possible de déposer ses données dans certains repositories pendant l’enquête et de convenir que les données ne seront ouvertes qu'une fois l’enquête terminée.
- Si vous choisissez d'utiliser un service tiers, vous devez veiller à ce que ce choix ne soit pas contraire aux politiques des financeurs, des institutions, des départements ou de groupes. Par exemple, prendre en compte la juridiction du pays dans lequel sont détenues les données ou la protection des données sensibles.

2.2) Accès et sécurité des données (1)

1. Sécurité sur les ordinateurs du chercheur ou membres de l'équipe de recherche

- Installez des sessions sécurisées sur les ordinateurs où sont stockées vos documents (de manière à ce que l'ordinateur en veille, il faille retaper les identifiants pour rouvrir la session).
- Dans l'idéal, utilisez des techniques de chiffrement pour vos fichiers de données.

2. Sécurité des échanges par courriels ou services de transfert de fichiers

De nombreux collègues s'échangent leurs données par courriel (en pièces-jointes d'e-mails) ou, pour les documents les plus volumineux, par services de transfert de fichiers en ligne.

- Les informations transférées via les réseaux peuvent être cryptées si nécessaire.
- Les informations confidentielles ne doivent pas être stockées sur des serveurs fournissant des services réseau (serveurs Web et de messagerie, par exemple). Les données confidentielles ne doivent être stockées que sur des ordinateurs non connectés à des réseaux.
- Plate-formes de transfert de fichiers volumineux : évitez absolument les FAI (Fournisseurs d'accès Internet) comme Free, Orange ou OVH ou We Transfer et privilégiez les outils open source : Framadrop, Droopy, FileX (système de transfert de fichier par interface web), FileZ (dépôt et gestion de fichier partagés grâce à une URL unique), BigFileSharing, MySecureShell.

2.2) Accès et sécurité des données (2)

3. Sécurité des dispositifs de stockage accessibles en ligne

- Si vous souhaitez stocker vos données en ligne de manière sécurisée, évitez les dispositifs du type Google Drive ou DropBox ou les différents services de cloud proposés par des entreprises. Privilégiez le stockage sur un serveur institutionnel sécurisé et pérenne dédié aux données de la recherche.
- Les personnes participant à l'enquête (chercheur.e(s), ingénieur.e(s), informaticien.ne(s), archiviste(s), documentaliste(s), etc.) doivent disposer de droits d'utilisation personnels pour lire et écrire des données (par exemple, noms d'utilisateur et mots de passe). Ceci est particulièrement important si les données de recherche peuvent être consultées via un réseau. Les mots de passe doivent présenter un degré élevé de sécurité, c'est-à-dire comprenant des chiffres, des lettres et des caractères spéciaux, majuscules et minuscules d'au moins 12 caractères. Donc le password avec votre nom, prénom, date de naissance, vous oubliez. Veillez également à renouveler régulièrement vos codes d'accès.
- Les droits d'accès doivent être définis pour tous les dossiers et fichiers, en particulier lorsqu'ils sont stockés sur un serveur plutôt que sur un seul ordinateur. Par exemple, il n'est pas nécessaire que tous les membres d'un projet de recherche aient le droit de modifier les fichiers de sauvegarde.
- Il convient de s'assurer que le système d'information n'enregistre aucun fichier temporaire ou autre résultant du traitement de données dans des dossiers accessibles à quiconque.

2.2) Accès et sécurité des données (3)

3. Sécurité des dispositifs de stockage accessibles en ligne

- Si vous souhaitez stocker vos données en ligne de manière sécurisée, évitez les dispositifs du type Google Drive ou DropBox ou les différents services de cloud proposés par des entreprises. Privilégiez le stockage sur un serveur institutionnel sécurisé et pérenne dédié aux données de la recherche.
- Les personnes participant à l'enquête (chercheur.e(s), ingénieur.e(s), informaticien.ne(s), archiviste(s), documentaliste(s), etc.) doivent disposer de droits d'utilisation personnels pour lire et écrire des données (par exemple, noms d'utilisateur et mots de passe). Ceci est particulièrement important si les données de recherche peuvent être consultées via un réseau. Les mots de passe doivent présenter un degré élevé de sécurité, c'est-à-dire comprenant des chiffres, des lettres et des caractères spéciaux, majuscules et minuscules d'au moins 12 caractères. Donc le password avec votre nom, prénom, date de naissance, vous oubliez. Veillez également à renouveler régulièrement vos codes d'accès.
- Les droits d'accès doivent être définis pour tous les dossiers et fichiers, en particulier lorsqu'ils sont stockés sur un serveur plutôt que sur un seul ordinateur. Par exemple, il n'est pas nécessaire que tous les membres d'un projet de recherche aient le droit de modifier les fichiers de sauvegarde.
- Il convient de s'assurer que le système d'information n'enregistre aucun fichier temporaire ou autre résultant du traitement de données dans des dossiers accessibles à quiconque.

2.2) Accès et sécurité des données (4)

4. Mises à jour des programmes, logiciels et formats.

- Les mises à jour pour les systèmes d'exploitation et les programmes doivent être installées aussi rapidement que possible. Il est recommandé d'utiliser un service de mise à jour qui installe automatiquement les mises à jour importantes et de garder à l'esprit que parfois les mises à jour logicielles peuvent causer des problèmes de compatibilité.
- Chaque document doit dans l’idéal être copié dans différents formats pérennes et sa lisibilité vérifiée régulièrement de manière automatique.

5. Protection antivirus

- Tous les ordinateurs utilisés dans le projet de recherche doivent disposer d’un logiciel antivirus à jour régulièrement et automatiquement installé.
- Vérifiez les informations que votre logiciel anti-virus collecte
- La plupart des virus sont conçus pour les OS propriétaires (Windows en particulier, de par sa position hégémonique sur le marché). Les ordinateurs sous Linux sont donc de fait beaucoup plus protégés de ce risque.

3) PARTAGE ET ARCHIVAGE

3.2) Favoriser le référencement des enquêtes : les identifiants numériques pérennes

Comment les personnes peuvent citer les données que vous avez produites/partagées lorsqu'ils les réutilisent ?

Dans cette perspective, l’attribution d’un **DOI** associé à l’enquête, voire à chaque document de l’enquête, facilite grandement la citation.

Voir notamment le dispositif DataCite

3.3) Mettre en place un dispositif de partage contrôlé des données à usage restreint

Une autre question concernant le partage est inverse : il ne s'agit plus de savoir quoi, comment et avec qui partager vos documents d'enquête mais si cela est possible. Autrement dit, existe-t-il des restrictions au partage des documents d'enquête, dues par exemple à la confidentialité, à l'absence de consentement ou aux droits de propriété intellectuelle par exemple.

Dans cette perspective, des dispositifs techniques existent pour permettre le partage contrôlé des données à usage restreint, ce qu'on appelle les *Virtual Data Enclave* (VDE).

À titre d'exemple, voir la VDE proposée par l'ICPSR

Pour plus de détails et de ressources, voir l'article issu de cette conférence qui devra être publié prochainement dans
Sociograph. Sociological Research Studies

**
Ce(tte) œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.**