L'éthique des données de la recherche en sciences humaines et sociales. Une introduction

Bernard Jacquemin, Joachim Schöpfel, Stéphane Chaudiron, and Éric Kergosien

Univ. Lille, EA 4073 - GERiiCO, F-59000 Lille, France.
{Prenom.Nom}@univ-lille.fr

Résumé

L'organisation de l'accès libre aux données scientifiques fait partie des objectifs de la recherche publique de la France. La volonté d'ouvrir les données de la recherche a été confirmée par le plan d'action national 2018-2020 dont l'engagement 18 vise à construire un écosystème de la science ouverte. Sur le terrain, la politique d'ouverture s'accompagne d'une forte incitation à mettre en œuvre des bonnes pratiques scientifiques compatibles avec certains principes définis au niveau européens comme « *FAIR Guiding Principles* » de la gestion et du pilotage des données de la recherche. Quelle est la dimension éthique d'une gestion « *FAIR* » des données de la recherche ? À partir d'une sélection de publications récentes, d'enquêtes, travaux et activités menées autour des données de la recherche, notre communication essaie de synthétiser plusieurs aspects de la dimension éthique de la gestion des données de la recherche, dans l'environnement français, dont la place de l'éthique dans les plans de gestion, les données personnelles, la crédibilité ou encore la sécurité des données.

Mots-clefs : plan de gestion, données personnelles, respect des personnes, crédibilité, sécurité, propriété intellectuelle, éthique, données de la recherche.

Abstract

The organisation of free access to scientific data is one of France's public research objectives. The commitment to open research data has been confirmed by the National Action Plan 2018-2020, whose commitment is to build an open science ecosystem. On the ground, the policy of openness is accompanied by a strong incentive to implement good scientific practices compatible with certain principles defined at European level as "FAIR Guiding Principles" for the management and steering of research data. What is the ethical dimension of "FAIR" research data management ? Based on a selection of recent publications, surveys, work and activities conducted around research data, our paper attempts to synthesize several aspects of the ethical dimension of research data management in the French environment, including the place of ethics in management plans, personal data, credibility and data security.

Keywords : data management plan, personal data, respect for individuals, credibility, security, intellectual property, ethics, research data.

1 Une gestion FAIR, mais aussi éthique ?

L'organisation de l'accès libre aux données scientifiques fait partie des objectifs de la recherche publique de la France (Code de la Recherche, article L112-1 alinéa e). La volonté d'ouvrir les données de la recherche a été confirmé par le plan d'action national 2018-2020 *Pour une action publique transparente et collaborative* dont l'engagement 18 vise à construire un écosystème de la science ouverte dans lequel « la science sera plus cumulative, plus fortement étayée par des données, plus transparente, plus intègre, plus rapide et d'accès plus universel (et qui) induit une démocratisation de l'accès aux savoirs, utile à la recherche, à la formation, à la société » (Etalab, 2018, 57). Ceci passera entre autre par l'incitation à une ouverture des données produites par les programmes de recherche publics, à partir de 2019.

Sur le terrain, la politique d'ouverture s'accompagne d'une forte incitation à mettre en œuvre des bonnes pratiques scientifiques compatibles avec certains principes définis au niveau européens comme *FAIR Guiding Principles* de la gestion et du pilotage des données de la recherche (Wilkinson *et al.*, 2016; European Commission, 2016) : les données doivent être faciles à (re)trouver, accessibles et si possible ouvertes, interopérables et réutilisables. Ces principes visent avant tout à faciliter le traitement automatique des données par des machines.

Quelle est la dimension éthique d'une gestion « FAIR » des données de la recherche ? Par rapport à l'accessibilité et l'ouverture des données, le Comité d'éthique du CNRS a mis en garde dès 2015 que « toutes ces consignes générales peuvent paraître en opposition avec les restrictions légales formulées au nom du respect de la vie privée, du droit d'auteur, de l'obligation de secret ou de la sécurité » (COMETS, 2015, 2). Les chercheurs eux-mêmes ont tendance à considérer l'éthique comme un « frein juridique » et principal facteur empêchant une libre diffusion de leurs données (Serres *et al.*, 2017).

Mais la gestion des données de la recherche mobilise aussi d'autres valeurs scientifiques, comme l'intégrité et la qualité des données, la transparence du processus, la solidarité, la réciprocité et l'engagement d'excellence vis-à-vis de la société, l'équité et le respect des personnes (Gundersen, 2016; Langat *et al.*, 2011). Aussi, dans l'environnement du traitement massif de données (*Big Data*), il est important de souligner que la dimension éthique ne se limite pas à la gestion des données mais intervient dès la collecte des données en amont et concerne l'ensemble du cycle de vie des données (Sula, 2016) et que la pratique scientifique, la législation et la politique de la recherche peuvent être en conflit en ce qui concerne la dimension éthique des données (Perry et Wilkinson, 2010).

Notre communication essaie de synthétiser plusieurs aspects de la dimension éthique de la gestion des données de la recherche, dans l'environnement français à partir d'une sélection de publications récentes, d'enquêtes, travaux et activités menées autour des données de la recherche par les professionnels de l'information et de la formation. Ainsi, nous avons identifié six facettes de cette relation entre l'éthique et les données de la recherche qui paraissent importantes notamment dans le domaine des sciences humaines et sociales (figure 1).

L'analyse de ces six facettes poursuit un double objectif, contribuer à une meilleure compréhension des données de la recherche sous l'aspect des sciences de l'informa-

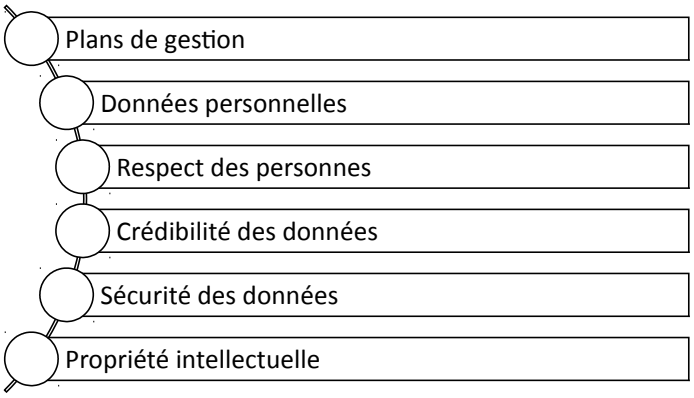

FIGURE 1 – Six facettes de la dimension éthique des données de la recherche

tion, et contribuer au développement de l'écosystème de la science ouverte à l'agenda la politique de recherche en France et en Europe. Les questions éthiques concernent l'ensemble des disciplines scientifiques. Néanmoins, notre étude se limitera essentiellement aux problèmes soulevés par la recherche en sciences humaines et sociales. De même, nous n'allons pas approfondir certains aspects juridiques et renvoyons les lecteurs à l'excellent guide d'analyse du cadre juridique en France publié par l'INRA (Becard *et al.*, 2017).

2 L'éthique comme élément des plans de gestion

L'engagement 18 du plan d'action national 2018-2020 prévoit la généralisation progressive de la mise en place de plans de gestion des données dans les appels à projets de recherche à partir de 2019. Ces plans de gestion (*data management plan* ou DMP), introduits notamment par le programme cadre Horizon 2020 (ou H2020) de la Commission européenne, sont devenus en quelques années indispensables dans les grands projets de recherche et, d'une manière plus générale, pour la bonne gestion des données collectées et produites.

Tous ces plans font le lien, directement ou indirectement, avec la dimension éthique. Ainsi, le premier modèle DMP du programme H2020, une trame simplifiée avec cinq questions, n'a pas de section « éthique » mais demande s'il y avait éventuellement des raisons éthiques (données personnelles ou autres) qui empêcheraient le partage et la libre diffusion des données. D'autres plans, comme ceux des Universités Paris Descartes et Paris Diderot, abordent également la question de l'éthique et en particulier de la protection des données à caractère personnel.

L'actuelle trame DMP du programme européen H2020 (*H2020 FAIR DMP*) intègre les principes de la gestion FAIR et évoque la dimension éthique à plusieurs reprises :

— Accessibilité : il faut justifier l'absence d'une libre diffusion de données par des raisons éthiques et/ou légales ;

— Réutilisation : la trame conseille pour la diffusion des données l'utilisation des licences proposées par EUDAT pour l'archivage à long terme [1] ;

— Sécurité : il faut décrire les mesures pour sécuriser les données, sans toutefois détailler la nature des données à protéger ;

— Aspects éthiques : la cinquième section du plan est consacrée à la dimension éthique et demande d'examiner s'il existe des problèmes éthiques ou juridiques pouvant avoir un impact sur le partage de données (consentement éclairé pour le partage des données et la conservation à long terme, etc.).

Cette cinquième section du plan *H2020 FAIR DMP* est facultative et sert avant tout à faire le lien avec la procédure de l'évaluation éthique (*ethics review*) du programme H2020 [2], laquelle contient un protocole détaillé concernant des questions de droits de l'homme et protection des êtres humains, de la protection des animaux, de la protection des données et de la vie privée, de la protection de l'environnement, de l'utilisation malveillante des résultats de la recherche, et de la conformité avec les lois internationales, européennes et nationales. Cette *ethics review* fait l'objet d'une analyse par une commission indépendante ; dans le cadre du plan de gestion, il ne s'agit donc pas de faire une autre analyse, redondante avec l'*ethics review*, mais d'expliciter comment la gestion des données tient compte des questions éthiques, par l'anonymisation des données personnelles, par les conditions et la durée de leur conservation, par le mode de diffusion, par le suivi par un comité d'éthique local, etc.

Pour résumer, les plans de gestion n'ont pas vocation à se substituer aux outils des comités d'éthiques. Leur fonction principale ici est de démontrer la prise de conscience de la dimension éthique par les chercheurs et de décrire, avec la terminologie de la gestion des données et par rapport aux principes FAIR, comment la gestion des données va intégrer cette dimension.

3 Les données à caractère personnel

Les données à caractère personnel, c'est-à-dire pour la CNIL « toute information relative à une personne physique susceptible d'être identifiée, directement ou indirectement » [3], cristallisent les problèmes et enjeux de la gestion des données dans un environnement du *Big Data*, par le clivage entre le potentiel immense de la science des données et les limites imposées par la réglementation juridique et éthique (Metcalf et Crawford, 2016). La politique d'ouverture des données par défaut se heurte nécessairement à l'obstacle majeur de la protection des données personnelles, notamment dans le domaine biomédical. Ainsi, J.-G. Ganascia, président du Comité d'éthique du CNRS, a attiré l'attention sur le fait que « les informations relatives à la santé, comme en produisent beaucoup les sciences médicales et parfois les sciences humaines, sont

1. https://eudat.eu/
2. http://ec.europa.eu/research/participants/docs/h2020-funding-guide/grants/from-evaluation-to-grant-signature/grant-preparation/ethics_review_en.htm
3. https://www.cnil.fr

sensibles parce que susceptibles d'aboutir à l'identification de personnes » en suggérant une approche disciplinaire afin de dresser une typologie des différentes catégories de données pour en préciser le statut et de garantir un partage raisonné et équitable [4].

Suivant la même logique, le Comité d'éthique du CNRS recommande dans un avis de 2018 « dans la mesure du possible l'ouverture de plates-formes de données en prenant en compte l'exigence de la protection des données personnelles et en fournissant les moyens de les traiter sans biais » – une approche compatible avec la politique de la Commission européenne pour l'ouverture des données, « *as open as possible, as closed as necessary* ». Par ailleurs, la Commission, dans ses dernières recommandations pour l'accès et la conservation de l'information scientifique du 25 avril 2018, évoque les données personnelles (*privacy*) comme seule raison éthique pouvant justifier un accès limité aux résultats de la recherche, au même titre que « *trade secrets, national security, legitimate commercial interests and (…) intellectual property rights of third parties* » [5].

La *Loi pour une République numérique* de 2016 avait stipulé que la réutilisation des données de la recherche était libre, sauf si ces données « sont protégées par un droit spécifique ou une réglementation particulière », ce qui est bien entendu le cas pour les données personnelles. Pour la gestion des données dans le cadre d'un projet, ces données personnelles posent plusieurs problèmes, dont :

- la conformité avec la loi *Informatique et Libertés* ;
- en particulier, l'autorisation du partage et de la publication des données ;
- le cas échéant (projets avec les hôpitaux etc.), la conformité avec la réglementation pour les données de la santé ;
- pour les projets internationaux, la conformité avec les différentes réglementations nationales (Hoeyer *et al.*, 2017; Dove et Garattini, 2018) ;
- et l'autorisation d'une réutilisation des données, notamment de données qualitatives (Grinyer, 2009).

Aucun de ces problèmes n'est vraiment nouveau. Mais dans le contexte de l'ouverture des données, le consentement du partage et de la réutilisation des données qui ont été auparavant collectées dans le cadre d'une recherche précise devient un enjeu essentiel. S'il semble légitime d'organiser, pour leur réutilisation, l'ouverture des données de recherche surtout quand elles ont été collectées à l'aide de financements publics, cette légitimité se heurte à un questionnement éthique : les données ont été collectées dans un cadre et dans un but précis, qui légitimait leur collecte auprès des individus ou organismes concernés. La réutilisation de ces données dans un autre cadre, dans un autre but, risque de rendre caduque cette légitimité, ce qui remet en cause le libre accès à ces données pour leur réutilisation. Nous y reviendrons plus loin.

Au moment de la rédaction du manuscrit (mai 2018), il est trop tôt pour évaluer l'impact du nouveau règlement européen sur la protection des données personnelles (RGPD). L'avis est partagé notamment concernant l'exploitation secondaire des données qui *a priori* n'a pas besoin d'un nouveau consentement à condition que la réutilisation ne soit pas incompatible avec l'objectif initial ou principal de la collecte des

4. http://cnrsinfo.cnrs.fr/intranet/actus/170615-jg-ganascia.html

5. https://ec.europa.eu/digital-single-market/en/news/recommendation-access-and-preservation-scientific-information

données. Mais en cas de litige, est-ce que les tribunaux vont appliquer une interprétation large ou plutôt étroite de ce consentement initial ? Même si le RGPD semble plus flexible pour la recherche scientifique, il n'est pas certain si son application va freiner ou accélérer l'ouverture des données de la recherche (Glinos, 2018; Lauber-Rönsberg, 2018). Il est certain, par contre, que le RGPD va élargir la notion des données de la santé par nature, du fait de leur croisement avec d'autres données et en raison de leur destination, et que par conséquent les droits des personnes seront mieux protégés ce qui rendra la publication et la réutilisation des données issues de la recherche médicale plus difficiles.

4 Respects des personnes, conflits d'intérêt

Au-delà de la nécessaire protection des informations à caractère personnel, c'est la légitimité même de ces données qui peut être questionnée d'un point de vue éthique. En effet, la participation d'un individu à une action de recherche relève d'une volonté personnelle, et touche donc à la liberté individuelle. Le chercheur a d'ailleurs le devoir d'informer les personnes qui participent à son activité de la nature de son travail et de la finalité de sa recherche pour pouvoir procéder au recueil d'informations. La pertinence des données collectées doit aussi se justifier au regard des objectifs de la démarche.

C'est dans les disciplines liées à la santé (recherche médicale, biologie…) que le cadrage éthique se fait sentir avec le plus d'acuité, notamment parce qu'il est porté institutionnellement et qu'il affecte le devenir du travail de recherche. Des comités d'éthique doivent en effet valider toute démarche impliquant des sujets humains avant qu'une expérimentation soit menée et ses résultats publiés. En France, un comité d'éthique national et des instances d'établissements peuvent être interpellés pour valider la légitimité des actions envisagées au regard des hypothèses émises et des visées de l'expérimentation, et cet avis porte également sur les données collectées elles-mêmes. Il sera par exemple nécessaire de justifier l'usage de techniques invasives (endoscopie, colorant circulant dans le système sanguin ou le système digestif) pour collecter des données. De même, la collecte de données à caractère sensible – comme l'appartenance à une religion ou l'origine ethnique – devra être encadrée strictement : l'exemple des recherche menées en Amérique sur les peuples autochtones montre que la pertinence de cette collecte doit être démontrée à la fois du côté des chercheurs et du côté des personnes sollicitées (Asselin et Basile, 2012; Quinless, 2018) [6]. En plus de l'avis favorable du comité d'éthique, les chercheurs sont tenus d'informer les sujets de leurs expérimentations à la fois du déroulement et de la finalité de l'expérimentation, et de leur donner accès aux données collectées : c'est le « consentement éclairé », formalisé par un document signé, que les individus peuvent annuler durant toute la durée de l'expérimentation. En sciences humaines et sociales, ce sont essentiellement les domaines proches de la santé également, tels la psychologie ou la sociologie de la santé, qui doivent en passer par ces comités et ces formulaires. Cependant, tant au niveau national qu'international, la tendance est à systématiser la sollicitation d'une

6. *Cf.* aussi le groupe de travail de l'organisation EUDAT sur la gestion des données dites « sensibles » https://www.eudat.eu/a-eudat-working-group-on-sensitive-data-management.

évaluation éthique pour « tous les projets de recherche qui requièrent la participation d'individus ou l'utilisation de données les concernant » (Orientations du Fonds québécois de la recherche sur la société et la culture, 2002, 13), ainsi qu'à établir le consentement éclairé des sujets avant la collecte de données.

Dès lors, la science ouverte et l'ouverture des données de la recherche sont confrontées à un écueil éthique de taille. En effet, la réorientation des objectifs de recherche, l'accès et la réutilisation des données de la recherche, même collectées selon des standards éthiques élevés, s'exposent à une sortie du cadrage défini originellement et présenté d'une part au comité d'éthique et de l'autre aux participants pour l'obtention de leur consentement éclairé. Or il apparaît que ce qui rend éthiquement légitime la participation active d'individus à un processus de recherche repose sur trois paramètres : un lien explicite entre la méthodologie de recherche et l'objectif de développement des connaissances visé, qui au regard de l'état des connaissances scientifiques actuelles rende cette démarche nécessaire pour atteindre cet objectif. Ce lien doit être évalué et validé par un comité d'éthique ; un accord de participation active à la démarche de recherche de la part des individus sollicités, formalisée par la signature d'un formulaire de consentement ; la communication d'une information explicitant à la fois l'expérimentation menée et les modalités d'accès et de protection des données collectées, également à travers le formulaire de consentement éclairé. C'est donc par la communication aux individus d'une information du devenir ultérieur des données de la recherche que passe leur ouverture.

Encore cette communication, pour garantir la protection des individus et l'expression de leur libre arbitre, doit-elle être exempte de tout conflit d'intérêt. En effet, aucune des parties impliquées dans la collecte, l'utilisation ou la réutilisation des données ne peut être suspectée d'en tirer avantage au détriment des autres parties. Si les comités d'éthique permettent en principe de se préserver de tels conflits d'intérêt lors de la collecte et de l'exploitation des données de recherche, il n'en va pas de même en cas de réutilisation. En effet, l'ouverture des données de la recherche est susceptible de laisser accès aux données pour un usage qui se ferait au détriment des individus, par exemple en cas de conservation d'informations liées à des comportements devenus illicites ou inappropriés. La protection des individus passe alors non seulement par un cadrage strict des conditions de réutilisation des données initiales, mais aussi par une suppression rigoureuse de toute information permettant de faire le lien avec les individus impliqués.

5 La crédibilité des données

Une quatrième manière d'aborder l'éthique des données de la recherche passe par la crédibilité, par la confiance (*trustworthiness*) qui peut être accordée aux services de données (entrepôts, plates-formes de gestion, archives, etc.). En effet, leur mise à disposition, en particulier sur des infrastructures techniques institutionnelles, pose la question de leur valeur intrinsèque et de leur équité au regard du poids symbolique qu'elles retirent de l'affichage institutionnel. Un jeu de données déposé sur une plate-forme officielle bénéficie en effet du capital-confiance accordé à l'institution, comme garant de qualité, et *a contrario* sa publication l'expose à l'examen des com-

munautés scientifiques ou non, qui peuvent en remettre en question la probité ou la pertinence, et en arriver à valider ou invalider des textes scientifiques pourtant publiés. On peut distinguer trois aspects différents, tous en lien avec la crédibilité des données (figure 2) :

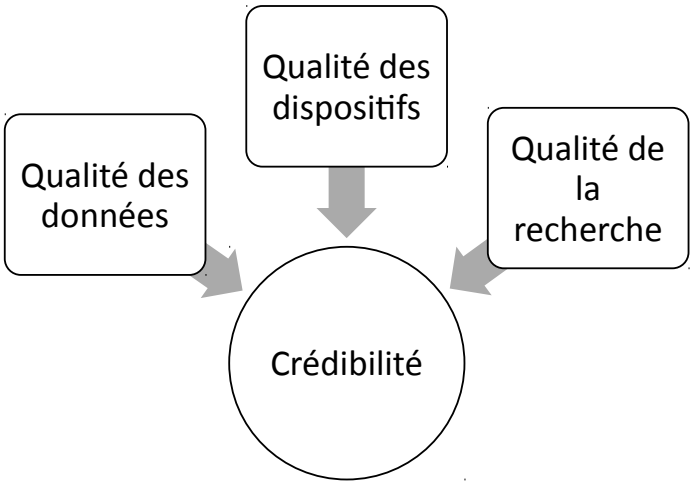

FIGURE 2 – La crédibilité des données de la recherche

La qualité des données : l'éthique des métiers de l'information souligne l'engagement de fournir une information de qualité, sans censure, sans discrimination, sans biais [7]. Dans le cadre de la gestion des données, ce terme recouvre plusieurs significations, aussi bien de la véracité (pertinence) de l'information que de la cohérence entre plusieurs sources (bases de données, etc.), la conformité avec l'usage (utilité, finalité), l'absence de problème d'accès, etc. Provoost (2015) a décrit la qualité des données sous l'angle de la réutilisation secondaire (utilisation non appropriée, non qualité, usage non éthique…). La question de la valeur et de la véracité des données est régulièrement discutée notamment par rapport au concept du *Big Data* dont elles constituent des éléments-clés (Teboul et Berthier, 2015). Du point de vue informatique, il s'agit d'une question technique. Mais du point de vue scientifique, la question technique se transforme en problème éthique : comment éviter d'exploiter et surtout de publier des données caduques, erronées, falsifiées, non pertinentes, inexploitables ?

La qualité des dispositifs : quelle est la sécurité des données ? Quelle est la garantie de conservation à long terme ? Quelle est la qualité de la description et documentation des données, par exemple le degré de normalisation des métadonnées ? Quel est le degré d'interopérabilité de la plate-forme de diffusion des données ? *A priori*, ces questions n'ont pas de lien direct avec l'éthique des données. Mais dans la mesure où toutes ces variables contribuent à la qualité et

7. *Cf.* le code éthique de la Fédération Internationale des Associations et Institutions de Bibliothèques (IFLA) https://www.ifla.org/faife/professional-codes-of-ethics-for-librarians.

crédibilité des données, ce lien existe quand-même, par rapport à la finalité de la diffusion, y compris la possibilité d'une réutilisation, et par rapport à la promesse de service du dispositif. Les démarches pour renforcer et promouvoir la qualité des dispositifs sont avant tout l'audit et la labellisation ou certification des entrepôts de données et archives numériques (*DINI*[8], *Data Seal of Approval*[9], *Core Trust Seal*[10], *Drambora*[11] etc.). Par ailleurs, quand il s'agit des données de la recherche, les système d'information de la recherche ont tendance à évaluer davantage la qualité des entrepôts et les modalités de diffusion que la qualité des données elles-mêmes (Schöpfel *et al.*, 2017).

La qualité de la recherche : la publication des données de la recherche est censée faciliter la reproductibilité des recherches, d'augmenter la transparence du travail scientifique, de prévenir la falsification des données et de contribuer ainsi à l'amélioration des résultats et des connaissances. Regardé de près, ce lien entre la diffusion des données et l'intégrité et la responsabilité du chercheur est double voire dialectique : d'une part, l'ouverture des données renforce l'intégrité et la valeur du travail scientifique ; d'autre part, la responsabilité et l'intégrité du chercheur (ou plutôt de l'équipe ou de la structure de recherche) contribuent à rendre les données déposées et diffusées crédibles et potentiellement réutilisables.

6 La sécurité des données

Comme pour tout dispositif de stockage et de conservation de données en ligne, la sécurisation des données est un enjeu de première importance. D'un point de vue technique, il s'agit à la fois de veiller à l'intégrité de la conservation des données, et de sécuriser leur accès en lecture et en écriture. La valeur des données conservées réside en effet dans leur intégrité, et il n'est pas acceptables qu'elles soient altérées, de quelque manière que ce soit : la perte, la modification, l'adjonction incontrôlée de toute valeur est susceptible non seulement de fausser une action de recherche menée à partir d'un jeu de données corrompues, mais également d'invalider une publication déjà diffusée si l'éditeur demande à avoir accès aux données qui la portent, comme cela commence à se produire.

Les accidents techniques – *crashes* disque, incompatibilités diverses : formats, encodages, logiciels, matériels…, erreurs de transmission, etc. – doivent bien sûr être pris en compte et surmontés pour offrir un accès équitable et fidèle aux sources scientifiques. Mais la problématique de la sécurisation des données porte également sur un large volet humain. Si des maladresses dans la gestion de ces données sont toujours possible, c'est plus particulièrement la malveillance, souvent massive sur les réseaux numériques qui opère ici, même (et peut-être surtout) lorsque l'objectif poursuivi est le développement des connaissances. L'encyclopédie Wikipédia, dont une partie non négligeable des contributions est l'œuvre tantôt de vandales et tantôt de prosélytes

8. *DINI certificat* https://edoc.hu-berlin.de/handle/18452/2148.
9. *Data Seal of Approval* https://www.datasealofapproval.org/en/.
10. *Core Trust Seal* https://www.coretrustseal.org/.
11. *Drambora* https://www.repositoryaudit.eu/about/.

(Foglia, 2009; Casilli, 2015), en est un exemple significatif, et seuls des mécanismes de protection tels que la vigilance participative (Cardon et Levrel, 2009) assurent un certain crédit à la ressource. Une protection à la fois des données déposées et du dispositif de conservation est donc absolument nécessaire. Bien plus, l'ouverture même des données peut le cas échéant nécessiter des restrictions, temporaires ou non, pour des raisons de confidentialité notamment, comme c'est le cas pour les archives. Il apparaît donc nécessaire de soumettre les données ouvertes à une contrainte, tant pour celui qui dépose que pour celui qui consulte : son identification devient nécessaire. Cela pose une nouvelle question éthique : dans quelle mesure l'identification des utilisateurs et la mise en lien de ces utilisateurs avec des jeux de données est-elle légitime ?

Enfin, assurer la sécurité des données, c'est encore faire en sorte que les utilisateurs soient en mesure de vérifier eux-mêmes la nature, la qualité, la pertinence et surtout l'intégrité des données auxquelles ils ont accès. Au niveau technique, l'utilisation d'une *somme de contrôle*, ou même d'un *code correcteur*, permet de déterminer automatiquement le niveau d'intégrité des données, voire de récupérer leur état initial, à condition que ces codes soient générés et conservés de bonne foi. Un descriptif précis et détaillé des jeux de données permet aux personnes intéressées de contrôler elles-mêmes leur intégrité sans passer par le truchement d'un dispositif technique, et de se fier à leur propre jugement, de décider du niveau de confiance qu'elles sont prêtes à leur accorder.

7 La propriété intellectuelle

Une dernière interrogation éthique porte sur la propriété intellectuelle attachée aux données de la recherche. Ce questionnement porte sur trois aspects considérés comme primordiaux par les acteurs de la recherche : la préservation à la fois de l'intégrité des données et de leur lien avec les responsables de leur collecte, dans un univers numérique dans lequel chacun semble faire son marché librement et sans respect pour la propriété ; la distinction entre propriété intellectuelle et propriété industrielle, souvent mal comprise ou interprétée, où l'éthique et la légalité ne semblent pas toujours aller de concert ; une dissolution du lien de paternité des données lorsque la recherche est financée sur fonds publics.

Sur les réseaux numériques, la perte de consistance physique des produits disponibles et la présence massive d'une information gratuitement accessible conduit la notion de propriété à se dissoudre dans cette disponibilité. Cet affaiblissement a pour conséquence une utilisation importante d'informations et de bribes de textes récupérées sur le web sans être rattachées à aucune source, et donc une perte du lien de paternité, même dans les milieux scientifiques et universitaires (Simonnot, 2014). La crainte est donc grande pour les chercheurs, en cas d'ouverture de leurs données, de les voir récupérées et utilisées pour des actions dont ils n'auraient pas forcément connaissance, et sans en retirer aucun bénéfice, pas même moral grâce à une citation. On notera toutefois qu'il pourrait en aller de même avec les textes de recherche publiés, et donc diffusés, sans pour autant que le monde de la recherche s'en inquiète particulièrement.

Par ailleurs, le rapport que les acteurs de la recherche entretiennent avec leurs données dépasse souvent la simple demande d'explicitation de paternité, et se rapproche plutôt du contrôle. Si l'ouverture des données leur pose question, la réutilisation des données en pose plus encore – qui, dans quel but, pour quels usages et quels bénéfices ? –, et même une simple consultation. Or les approches éthiques de l'information (Zimmermann et Foray, 2001; IFLA, 2012) considèrent que les produits « immatériels » sont à la fois non excluables (la transmission d'une information la rend propre à être réutilisée, à condition de procéder à une indication de paternité) et non rivales (la transmission d'une information à un tiers ne diminue ni la qualité ni la quantité de l'information disponible). L'ouverture des données ne pose donc ici aucun problème éthique en cas d'attribution de paternité. C'est au niveau juridique qu'il n'en va pas de même : la loi rend excluable l'information pour une durée limitée (par exemple le droit d'auteur, éteint en France 70 ans après le décès du dernier auteur vivant). Encore cette propriété, de nature industrielle, doit-elle être correctement identifiée. Elle n'est que rarement l'exclusivité du chercheur, rémunéré pour son activité de recherche, et qui ne peut donc décider seul du devenir des données produites.

Cela nous amène au troisième aspect de la propriété des données. Une grande partie des actions de recherche menées en sciences humaines et sociales sont pris en charge, en tout ou en partie, par des fonds publics, que ce soit via le salaire des chercheurs ou à travers le financement de projets par une institution publique. D'un point de vue strictement légal, le financeur a donc un avis prépondérant quant au sort réservé aux données. Au niveau éthique, la question posée concerne l'accès par les citoyens aux données et résultats produits par une recherche financée par de l'argent public : les milieux favorables à la mouvance *Open* (logiciel libre, science ouverte, revues en libre accès, archives ouvertes) se sont prononcés, suivis timidement par les pouvoirs publics [12], pour cette interprétation.

8 Conclusion

La dimension éthique de la gestion des données de la recherche engage aussi bien les chercheurs que les institutions. Ainsi, le Comité d'éthique du CNRS recommande que les chercheurs « doivent prendre conscience de leur responsabilité individuelle, déontologique et éthique, vis à vis de la communauté à laquelle ils appartiennent » et qu'ils participent à définition de bonnes pratiques ; quant aux institutions, le Comité souligne l'importance d'informer les chercheurs « sur les limites de ce partage », notamment par rapport aux données à caractère personnel, et de « recenser les obstacles au partage éthique des données (propriété intellectuelle des données et statut *sui generis* des banques de données), afin de promouvoir des communs scientifiques et d'ériger les données de la science en données d'intérêt général » (COMETS, 2015 : 8-10).

D'après plusieurs enquêtes récentes, les chercheurs sont généralement conscients de la déontologie comme enjeu pour le partage des données (Serres *et al.*, 2017). Ils n'ont pas nécessairement une connaissance précise des recommandations sous forme

12. Par exemple avec la licence Etalab (décret 2017-638).

de règles éthiques ou incitations politiques mais n'expriment pas non plus – à l'exception des doctorants – un besoin prioritaire d'assistance ou d'aide dans ce domaine (Prost et Schöpfel, 2015). Ce constat est tout à fait cohérent avec les directeurs des unités de recherche du CNRS qui considèrent majoritairement que les chercheurs de leur laboratoire sont plutôts compétents dans ce domaine (Schöpfel *et al.*, 2018).

L'éthique n'est peut-être pas l'aspect le plus important quand il s'agit de la gestion des données de la recherche et de la mise en place d'un écosystème de la science ouverte. Néanmoins il s'agit d'un aspect indissociable et incontournable de la recherche scientifique en général et de la gestion des données en particulier. À la place de penser l'éthique comme un obstacle à la libre diffusion des résultats de la recherche, il faudrait (re)positionner la dimension éthique au cœur de la gestion des données, comme un critère et une condition nécessaire à la bonne pratique scientifique (Crow *et al.*, 2006 pour le lien entre consentement éclairé et qualité des résultats). En guise de conclusion, trois suggestions qui vont dans ce sens :

1. Renforcer le lien entre plans de gestion et protocoles éthiques : rendre obligatoire l'évaluation de la dimension éthique dans les plans de gestion, lors de leur généralisation progressive à partir de 2019, et formaliser le lien entre les deux documents par des identifiants et/ou adresses pérennes.
2. Traduire la dimension éthique dans les licences d'accès aux données : enrichir les métadonnées d'une manière précise avec les droits de réutilisation, avec une information détaillée et documentée sur l'objet et la portée des consentements de la part des personnes (sujets) concernés, et ce notamment dans le contexte du RGPD (ce qui peut inclure le développement de « licences éthiques » similaires à la gestion des licences *Creative Commons*, Lewis *et al.*, 2017). Pour les formulaires de consentement, cela voudrait dire : utiliser un vocabulaire non-restrictif, standardisé et clair ; et informer les participants des pratiques de partage de données, et des limites que celles-ci imposent à l'anonymat, à la confidentialité et à la possibilité de retrait.
3. (Re)définir le rôle et les compétences des comités d'éthique : créer des comités d'éthique sur les données de la recherche par discipline ou établissement (préconisation par le Comité d'éthique du CNRS), et éviter des redondances, incohérences et inefficiences en cas de projets multipartenaires ou internationaux (Dove et Garattini, 2018).

Pour le travail des comités d'éthique dans le domaine biomédical, Thorogood et Knoppers (2017) ont proposé plusieurs fonctions :

— évaluer les plans de partage afin de vérifier qu'ils assurent correctement la protection de la vie privée, l'accès aux données dans un délai raisonnable et le suivi de l'utilisation pour fins de recherche ;

— s'assurer que les formulaires de consentement permettent un accès suffisamment large aux données et la réutilisation de celles-ci pour permettre le partage ;

— se coordonner avec les comités d'accès et les autres comités d'éthique pour éviter les lacunes ou la duplication inutile de la surveillance ;

— renforcer le partage de données, de concert avec les institutions, les revues scientifiques et les organismes réglementaires.

Adapter ces propositions aux particularités des sciences humaines et sociales, renforcer le rôle du comité d'éthique sur le campus universitaire pourrait être une manière de réconcilier la politique d'ouverture des données et les réglementations éthique en vigueur, y compris le RGPD, et de contribuer ainsi à la mise en place d'un écosystème de la science ouverte et éthique.

Remerciements

Une partie des travaux à l'origine de cette communication a été réalisée dans le projet *D4Humanities* (*Deposit of Dissertation Data in Social Sciences and Humanities – A project in Digital Humanities*). Ce projet est financé dans le cadre des projets structurants de la MESHS 2017-2018 (Contrat de plan État-Région « ISI-MESHS »), par la MESHS et le Conseil Régional Hauts-de-France.

Références

Asselin, H. et Basile, S. (2012). Éthique de la recherche avec les peuples autochtones. Qu'en pensent les principaux intéressés ? *Éthique publique. Revue internationale d'éthique sociétale et gouvernementale*, 14(1).

Becard, N., Castets-Renard, C., Chassang, G., Dantant, M., Freyt-Caffin, L., Gandon, N., Martin, C., Martelletti, A., Mendoza-Caminade, A., Morcrette, N. et Neirac, C. (2017). Ouverture des données de la recherche. Guide d'analyse du cadre juridique en France. Rapport technique DOI : 10.15454/1.481273124091092E12, INRA, Paris.

Cardon, D. et Levrel, J. (2009). La vigilance participative. Une interprétation de la gouvernance de Wikipédia. *Réseaux*, 154(2):53–89.

Casilli, A. A. (2015). Le wikipédien, le chercheur et le vandale. *In* Barbe, L., Schafer, V. et Merzeau, L., éditeurs : *Wikipédia, objet scientifique non identifié*, pages 91–103. Presses Universitaires de Paris Ouest, Paris.

COMETS (2015). Les enjeux éthiques du partage des données scientifiques. Rapport technique, Comité d'éthique du CNRS, Paris.

Crow, G., Wiles, R., Heath, S. et Charles, V. (2006). Research Ethics and Data Quality : The Implications of Informed Consent. *International Journal of Social Research Methodology*, 9(2):83–95.

Dove, E. S. et Garattini, C. (2018). Expert perspectives on ethics review of international data-intensive research : Working towards mutual recognition. *Research Ethics*, 14(1):1–25.

Etalab (2018). Pour une action publique transparente et collaborative : plan d'action national pour la France 2018-2020. Rapport technique, Secrétariat d'État chargé de la Réforme de l'État et de la Simplification, Paris.

European Commission (2016). H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020. Version 3.0. Rapport technique, European Commission Directorate-General for Research & Innovation, Bruxelles.

Foglia, M. (2009). Faut-il avoir peur de Wikipédia ? *Études*, 410(4):463–472.

Glinos, K. (2018). Global data meet EU rules. *Science*, 360(6388):467.

Grinyer, A. (2009). The ethics of the secondary analysis and further use of qualitative data. *Social Research Update*, 56:1–4.

Gundersen, L. C. (2016). Embedding Scientific Integrity and Ethics into the Scientific Process and Research Data Lifecycle. *In American Geophysical Union Fall Meeting Abstracts*, pages PA12B–05, San Francisco.

Hoeyer, K., Tupasela, A. et Rasmussen, M. B. (2017). Ethics Policies and Ethics Work in Cross-national Genetic Research and Data Sharing : Flows, Nonflows, and Overflows. *Science, Technology, & Human Values*, 42(3):381–404.

IFLA (2012). Code d'éthique de l'IFLA pour la bibliothécaires et les autres professionnel(le)s de l'information. Rapport technique, Fédération internationale des associations et institutions de bibliothèques, La Haye.

Langat, P., Pisartchik, D., Silva, D., Bernard, C., Olsen, K., Smith, M., Sahni, S. et Upshur, R. (2011). Is There a Duty to Share ? Ethics of Sharing Research Data in the Context of Public Health Emergencies. *Public Health Ethics*, 4(1):4–11.

Lauber-Rönsberg, A. (2018). Data Protection Laws, Research Ethics and Social Sciences. *In* Dobrick, F. M., Fischer, J. et Hagen, L. M., éditeurs : *Research Ethics in the Digital Age : Ethics for the Social Sciences and Humanities in Times of Mediatization and Digitization*, pages 29–44. Springer Fachmedien Wiesbaden, Wiesbaden.

Lewis, D., Moorkens, J. et Fatema, K. (2017). Integrating the Management of Personal Data Protection and Open Science with Research Ethics. *In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 60–65, Valencia, Spain. Association for Computational Linguistics.

Metcalf, J. et Crawford, K. (2016). Where are human subjects in Big Data research ? The emerging ethics divide. *Big Data & Society*, 3(1):2053951716650211.

Orientations du Fonds québécois de la recherche sur la société et la culture (2002). *Éthique de la recherche sociale. Consentement libre et éclairé. Confidentialité et vie privée*. Fonds de recherche sur la société et la culture, Québec.

Perry, M. et Wilkinson, M. A. (2010). The Creation of University Intellectual Property : Confidential Information, Data Protection, and Research Ethics. *Canadian Intellectual Property Review*, 26:93–122.

Prost, H. et Schöpfel, J. (2015). Les données de la recherche en SHS. Une enquête à l'Université de Lille 3. Rapport final. Rapport technique, Université de Lille 3, Villeneuve d'Ascq.

Provoost, V. (2015). Secondary use of empirical research data in medical ethics papers on gamete donation : forms of use and pitfalls. *Monash Bioethics Review*, 33(1):64–77.

Quinless, J. (2018). Surveying the Landscape : Research Data Management, Data Governance and Ethics. British Columbia Research Libraries Group Lecture Series.

Schöpfel, J., Ferrant, C., André, F. et Fabre, R. (2018). Research data management in the French National Research Center (CNRS). *Data Technologies and Applications*, 52(2):248–265.

Schöpfel, J., Prost, H. et Rebouillat, V. (2017). Research Data in Current Research Information Systems. *Procedia Computer Science*, 106:305–320.

Serres, A., Malingre, M.-L., Mignon, M., Pierre, C. et Collet, D. (2017). Practices, mental images and expectations of researchers about research data in the humanities : a survey at Rennes 2 University. Research Report, Université Rennes 2.

Simonnot, B. (2014). Le plagiat universitaire, seulement une question d'éthique ? *Questions de communication*, 26:219–233.

Sula, C. A. (2016). Research Ethics in an Age of Big Data. *Bulletin of the Association for Information Science and Technology*, 42(2):17–21.

Teboul, B. et Berthier, T. (2015). Valeur et Véracité de la donnée : enjeux pour l'entreprise et défis pour le Data Scientist. *In Actes du colloque "La donnée n'est pas donnée"*, Paris. École Militaire.

Thorogood, A. et Knoppers, B. M. (2017). Can research ethics committees enable clinical trial data sharing ? *Ethics, Medicine and Public Health*, 3(1):56–63.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Santos, L. B. d. S., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., Hoen, P. A. C. t., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., Schaik, R. v., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Lei, J. v. d., Mulligen, E. v., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. et Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship.

Zimmermann, J.-B. et Foray, D. (2001). L'économie du logiciel libre. Organisation coopérative et incitation à l'innovation. *Revue économique*, 52(hors-série):77–93.