AVIS n° 2015-30

LES ENJEUX ÉTHIQUES DU PARTAGE DES DONNÉES SCIENTIFIQUES

Avis n° 2015-30 approuvé en séance plénière du COMETS le 7 mai 2015

RESUMÉ

Le développement massif d'outils informatiques de collecte, de mesure et de traitement a changé le rôle des données dans la production du travail scientifique. Le mouvement de partage des données scientifiques (*data sharing*) consacré par les dispositifs internationaux comme la Déclaration de Berlin en 2003 est une réponse au besoin d'échanger le plus rapidement possible les résultats obtenus et de surmonter les obstacles juridiques et techniques à la circulation de ces données. De même les politiques gouvernementales et européennes d'ouverture des données (*open data)* visent depuis quelques années à diffuser largement les données acquises grâce à des fonds publics. Cependant toutes les communautés scientifiques n'ont pas les mêmes contraintes vis à vis de cette ouverture. De même 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. Face à la complexité des obligations que rencontrent les chercheurs, cet avis a pour objet de réaffirmer le partage raisonné des données et d'inclure les nouvelles exigences de mise à disposition des données dans l'évaluation de leur travail. La question des données, qu'il s'agisse des verrous à surmonter comme des limites à leur ouverture est devenue une question cruciale dans la définition des politiques scientifiques.

SOMMAIRE

I. AUTO-SAISINE

Les données ont un rôle central dans la production scientifique de toutes les disciplines. Les chercheurs ont de plus en plus besoin de disposer de grandes masses de données, à côté de données de dimensions plus modestes, pour explorer, visualiser et comparer des résultats, valider des hypothèses ou en formuler de nouvelles, voire pour construire automatiquement de nouvelles connaissances par apprentissage machine. De grandes infrastructures et des plateformes communes sont créées en continu pour l'archivage, le stockage ou le traitement de l'information, bénéficiant des avancées récentes des technologies numériques. Les mouvements favorisant l'accès ouvert deviennent donc cruciaux. Les laboratoires de recherche publics et même privés ont de plus en plus souvent besoin de s'allier pour coordonner leurs efforts et réutiliser les données acquises par d'autres.[1]

Les attitudes vis-à-vis du partage et de l'ouverture sont très différentes suivant les types de données et les disciplines. En astrophysique ou en génomique par exemple, les rapprochements et comparaisons de données sont à l'évidence sources de nouvelles découvertes ; toute entrave à la circulation des résultats est contraire aux principes fondamentaux de mise en commun généralisée de ces connaissances. Pour d'autres disciplines, particulièrement en sciences humaines ou lorsque des enjeux industriels importants apparaissent, les données sont souvent collectées individuellement ou dans des conditions de circulation restreinte ; liées à l'objet de la recherche, elles ne peuvent être partagées qu'avec le même embargo que celui de la publication des résultats. Ajoutons que les données brutes n'ont pas toutes vocation à être conservées une fois qu'elles ont été traitées, car elles peuvent s'avérer trop volumineuses.

Le mouvement de partage des données scientifiques (*data sharing*) doit s'ajuster aux politiques gouvernementales plus récentes d'ouverture des données (*open data)* visant depuis quelques années à diffuser largement les données acquises grâce à des fonds publics, avec des objectifs et des contraintes juridiques et éthiques sensiblement différents. Cependant données publiques et données scientifiques se recoupent partiellement. Le Programme européen HORIZON 20202 consacre aussi le principe du libre accès aux publications et aux données scientifiques. 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. L'incitation à la valorisation de la recherche ou le devoir de respecter la confidentialité peut aussi s'opposer à la diffusion des données. Si un grand nombre de chercheurs sont favorables au principe d'ouverture des données, beaucoup se sentent démunis devant des contraintes qui peuvent paraître contradictoires. Le présent avis a pour objectif d'informer les chercheurs sur leurs obligations et de la portée de leurs choix vis à vis des données qu'ils collectent, partagent ou réutilisent, et de suggérer quelle doit être la réponse des établissements scientifiques à ces nouvelles obligations.

---

1 Voir l'article "Dix laboratoires mondiaux partageront données et chercheurs" (Le Monde 4 février 2014). Ce projet orchestré par le NIH demande notamment aux laboratoires privés ET publics de "ne pas développer leur propre médicament à partir des découvertes obtenues avant qu'elles n'aient été rendues publiques".

2 Il consacre le principe du «libre accès aux publications et aux données de la recherche» : voir http://www.horizon2020.gouv.fr

II. ANALYSE

A. Nature des données et contexte stratégique d'ouverture

Les données scientifiques considérées ici concernent toutes les données collectées dans le contexte de la recherche scientifique[3], c'est-à-dire les *données primaires* (empiriques, observées, mesurées) dont certaines n'ont pas vocation à être stockées et *a fortiori* à être partagées ; les *données secondaires*, dérivées des données primaires, annotées, enrichies, interprétées ajoutant de la valeur aux données initiales et pouvant impliquer d'autres acteurs ; les *métadonnées* qui structurent, gèrent, facilitent l'accessibilité des données primaires et secondaires et informent sur les conditions de partage. Ces données peuvent être des flux numériques issus de capteurs, ou des documents textuels, graphiques, picturaux, multimédia. L'écart entre le statut des données et celui des publications tend d'ailleurs à se réduire avec le concept de *science ouverte* qui consiste à diffuser les données et connaissances utilisées et construites au cours du processus d'élaboration et d'écriture de la publication scientifique[4].

Des accords et chartes successives ont jalonné l'histoire du mouvement du partage des données de la science. En 1996 pour la première fois des chercheurs impliqués dans le séquençage du génome humain, signent un ensemble d'accords consacrant les bases du partage des données dès leur production5. Puis la première définition de l'Open data est donnée par la Déclaration internationale sur le libre accès de Budapest le 14 février 2002, connue sous l'acronyme BOAI (Budapest Open Access Initiative)6. Notons que la clarification du sens des différents termes qui apparaissent ici ne s'est établie que progressivement. Aujourd'hui le terme d'Open data est réservé à l'ouverture des données obtenues sur fonds publics en général. On réserve en général le terme Data sharing au mouvement issu des communautés de chercheurs concernés par l'ouverture des données de la recherche sur lequel se concentre le présent avis. De nombreuses autres initiatives suivent, comme la Déclaration de Berlin de 2003 sur le libre accès à la connaissance dans toutes les sciences y compris les sciences humaines7 renforcée en 2005. La plupart des organismes scientifiques dont le CNRS ont signé ces déclarations et légitimé cette culture de l'accès ouvert8. Aujourd'hui la nécessité d'une réflexion sur l'ouverture et le partage des connaissances est inscrite dans le plan stratégique de la DIST du CNRS de 20149. Ajoutons

---

[3] H.Tjalsma & J. Rombouts, *Selection of research data- Guidelines for appraising and selecting research data,* The Hague: Data archiving and Networked services 2011, p.13,14 at http://www.dans.knaw.nl

[4] P.Uhlir «Revolution et evolution in scientific Communication: Moving from Restricted dissemination of publicly-funded knowledge to open knowledge environments»: http://www.communia-project.eu/communiafiles/Conf 2009_P_Uhlir_BS.pdf, 2008

[5] Bermuda principles, 1996, http://web.ornl.go/sci/techresources/Human_Genome/research/bermuda.shtml#1; Toronto Int[al] Data ReleaseWorkshopAuthors, 2009 *"Prepublication data sharing"*, *Nature* 461:168-170- http://www.nature.com/nature/journal/v461/n7261/full/461168a.html ; Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. *Data sharing in genomics-re-shaping scientific practice*. Nat Rev Genet. 2009 Vol 10, n°5, p.331-335

[6] http://www.budapestopenaccessinitiative.org/

[7] http://openaccess.inist.fr/?Declaration-de-Berlin-sur-le-Libre

8 En février 2012, le texte de la Déclaration de Berlin avait été signé par plus de 360 universités ou assimilées.

9 «Mieux partager les connaissances, une stratégie ouverte pour une information scientifique et technique d'avenir» : http://www.cnrs.fr/dist/z-outils/documents/STRATEGIE.pdf

que le libre accès aux données scientifiques fait désormais partie des objectifs de la recherche publique[10], avec cette précision : l'accès ne s'étend pas à la diffusion des données confidentielles, qui relèvent d'autres régimes juridiques. En effet certaines des données utilisées par les chercheurs pour conduire leurs études leur sont communiquées par des tiers sous réserve de confidentialité, en application d'une clause contractuelle ou d'une réglementation spécifique. Ces conditions peuvent être très restrictives, comme dans le cas des données fiscales, ouvertes à la recherche par la loi du 22 juillet 201311.

B. Très forte expansion de la masse des données, diversification de leur usage pour la recherche

L'activité scientifique s'appuie de plus en plus sur la création et l'utilisation partagées d'infrastructures de données multi-sources et multi-usages liées à trois types de changements : le perfectionnement des instruments de mesure et des capteurs de données brutes, les capacités informatiques considérablement accrues pour le stockage et l'archivage, l'internet collaboratif et la mise en réseau qui permettent un enrichissement direct en ligne de bases de données et de plateformes par de nombreux intervenants.

Il y a aujourd'hui une forte déperdition des données recueillies par les expérimentateurs. On estime que les publications permettent d'accéder à environ 10 % de celles-ci, le reste restant disponible mais non utilisé sur les disques durs d'ordinateurs. Dans certaines disciplines, des résultats valables et importants restent non publiés et beaucoup de données sont sous-utilisées ou perdues[12] (c'est en particulier le cas des données issues de résultats négatifs qui sont oubliées). Pour celles qui sont collectées par les grands instruments, les données brutes recueillies sont si massives qu'elles sont traitées directement en ligne sans être stockées, comme par exemple celles fournies par des observations spatiales. Il convient alors d'indiquer l'origine des données construites à partir de ces données brutes, même si celles-ci ont disparu. Il importe aussi de reconnaître la valeur du travail des personnels (chercheurs, ingénieurs, techniciens) qui ont contribué au traitement des données brutes pour les fournir sous une forme exploitable au reste de leur communauté. Ce travail souvent ingrat n'est pas toujours valorisé à la mesure de l'effort considérable qu'il représente en général.

Les grandes masses de données peuvent améliorer notre compréhension et aider à la prédiction des phénomènes en recourant à des techniques du type apprentissage-machine dans tous les domaines scientifiques, en particulier dans celui de la santé13.

---

10 Code de la Recherche, version en vigueur au 22 juillet 2013, article L112-1, alinéa «e» : http://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=4B5C99194EDEE3FB2A678D438F75DFC4.tpdila08v_2?idArticle=LEGIARTI000027747800&cidTexte=LEGITEXT000006071190&categorieLien=id&dateTexte=20150402

11 L'opportunité et l'ampleur de ces restrictions ne seront pas discutées. Nous les considérerons comme des éléments à prendre en considération pour délimiter le périmètre de ce qui est partageable.

12 Au début de 2014, cinq articles ont été publiés dans le Lancet sur le thème Research: increasing value, reducing waste. Voir notamment : An-Wen Chan, Fujian Song, Andrew Vickers, Tom Jefferson, Kay Dickersin, Peter C Gøtzsche, Harlan M Krumholz, Davina Ghersi, H Bart van der Worp, "Increasing value and reducing waste: addressing inaccessible research" in Lancet 2014; 383: 257–66.

13 http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-andepidemiology/WTDV030690.htm

Cependant le traitement de masses de données, du fait des volumes immenses d'information en jeu et, surtout, des dimensions de l'espace de descriptions, n'engendre pas des relations de causalité, mais des corrélations, qui suggèrent l'existence de causalités, sans le prouver. De nouvelles pratiques se développent, aboutissant à une véritable révolution épistémologique souvent désignée comme «data driven research»14 : partant de masses de données déjà constituées, des algorithmes explorent automatiquement des champs d'hypothèses et détectent des irrégularités ou des phénomènes imprévus qui échappent aux lois connues. Ces pratiques se développent, notamment dans des domaines comme les sciences de la Terre (sismologie) ou la biologie à grande échelle (génomique).

Un autre intérêt de la disponibilité de grandes masses de données, obtenue grâce aux fantastiques possibilités d'internet, est de permettre l'accès à la genèse et aux différentes étapes de l'élaboration d'une recherche, en montrant des résultats partiels interdépendants et évolutifs.

Les interactions cognitives se manifestent dans ce qu'on appelle les «*knowledge hubs*»[15] où coexistent plusieurs couches de connaissance plus ou moins élaborées, fruit dynamique d'un échange continu entre chercheurs. Toutes ces évolutions ont un effet heuristique sur le processus scientifique traditionnel.

L'accès aux données primaires devient alors déterminant pour permettre de vérifier leur qualité et aussi de juger de la méthodologie et de l'interprétation qui en découlent. La question du traçage et donc de la preuve scientifique devra aussi être interrogée à l'aune de ces nouvelles pratiques.

C. L'ouverture des données publiques et le partage des données scientifiques

A la différence du mouvement de partage de données développé par les chercheurs eux-mêmes, les politiques publiques d'*Open data* ont pris naissance hors de la communauté scientifique. En Europe, suite à la Directive sur la réutilisation des informations du secteur public[16], puis à la Directive créant des infrastructures d'information géographique[17], la majorité des pays ont adopté une politique visant à promouvoir l'ouverture des données publiques. Les Etats sont d'importants pourvoyeurs de données produites, reproduites, collectées, diffusées ou rediffusées par les administrations publiques dans le cadre de leurs missions institutionnelles. Il s'agit notamment de données démographiques, géographiques, météorologiques, économiques, financières, culturelles, touristiques, etc., qui visent à assurer la qualité et la continuité du service public. En France ETALAB[18], service qui gère

Voir aussi : M.J. Khoury & J.P.A. Ioannidis, «Big data meets public health» in Science, 26 November 2014, vol. 346, 6213 p.1054-1055

14 Leonelli S. Editorial: Making sense of data-driven research in the biological and biomedical sciences. Special Issue of Studies in the History and the Philosophy of the Biological and Biomedical Sciences: Part C 43:1, 2012, 1-3

15 H.D. Evers «Knowledge hubs and Knowledge clusters : designing a knowledge architecture for development» http://mpra.ub.uni-muenchen.de/8778/1/MPRA_paper_8778.pdf 2008

16 Directive 2003/98/CE du 17 novembre 2003 sur la réutilisation des informations du secteur public.

17 Directive 2007/2/CE du 14 mars 2007 établissant une infrastructure d'information géographique dans la Communauté européenne (appelée Directive Inspire).

[18]https://www.etalab.gouv.fr

l'*Open data* public sous l'autorité du Premier ministre, a pour mission de communiquer les données subventionnées sur fonds public avec mise à disposition libre et (quasi) gratuite, ce qui implique la possibilité de réutilisation de façon la moins contraignante possible. Un des autres objectifs de l'*Open data* est de permettre la valorisation, voire la monétisation de ces données en créant de la richesse pour les entreprises qui les exploitent. Enfin, un autre aspect plus collaboratif concerne les données mises à disposition pour les communautés de citoyens ou pour la société civile. Ces politiques ont en commun de créer des gisements de données accessibles et partageables et aussi de favoriser la transparence des modes de production de connaissances. Dans le cadre de *l'Open data* public, l'incitation est d'origine normative et s'applique légalement à l'ensemble des agents publics, y compris aux agents travaillant dans le service public de la recherche.

Les politiques qui promeuvent l'ouverture de données publiques n'ont pas les mêmes objectifs que celles du partage des données scientifiques. Pour éclaircir les régimes applicables, il serait utile de différencier données scientifiques et données publiques. Les données scientifiques produites sur des fonds publics ont, sauf exception[19], vocation à devenir publiques. Les données publiques ont vocation à devenir scientifiques lorsqu'elles concernent l'environnement, le climat, l'état de la société ou la santé. Les chercheurs doivent donc bénéficier de l'Open Data promu par l'Etat. Un exemple est fourni par l'assurance maladie qui dispose de la plus grande base de données du monde sur la santé : le SNIIRAM, alimenté en France depuis des décennies par les informations générées par la prise en charge de la totalité des consommations de soins et hospitalisations en France (20 milliards de lignes de prestations)[20]. Répondant à la demande relative à ce gisement de données, la Caisse Nationale d'Assurance Maladie entend développer sa politique d'ouverture, avec pour critères l'intérêt public (notamment pour la recherche), en fonction de la qualité du protocole, du besoin d'accéder aux données, de la sécurité des procédures et de la qualité du demandeur.

D. Les contraintes concernant le traitement de données personnelles

L'exemple précédent illustre bien les contraintes juridiques qui encadrent l'ouverture des données publiques. Les données de santé sont sensibles car susceptibles d'aboutir à l'identification des personnes. Le modèle de traitement de données qui figurait dans la loi dite Informatique et Libertés du 6 janvier 1978, modifiée en 200421 ne correspond plus aux modalités de traitement algorithmique actuel. Les dispositifs de la loi ne sont donc plus adaptés aux nouveaux contextes des données massives comme le signalent les chercheurs interrogés[22]. En effet la circulation rapide et ouverte des données entre chercheurs bouleverse l'ordre des procédures et rend les flux de données relativement autonomes par rapport à leurs sources ou leurs auteurs. Il devient souvent impossible de respecter le

[19] Par exemple certaines données de l'INSE ou de l'IGN

[20] Système National d'information inter-régimes de l'Assurance maladie. Voir : la proposition d'ouverture et de partage de ces «données publiques» : *Rapport sur la gouvernance et l'utilisation des données de santé* par Louis Bras et André Loth, Septembre 2013 p. 41- http://www.drees.sante.gouv.fr/IMG/pdf/rapport-donnees-de-sante-2013.pdf. La loi relative à la santé (N° 2302) votée en avril 2015 consacre l'open data en santé dans son article 47 en créant «les conditions d'un accès ouvert et sécurisé aux données de santé».

[21] Voir : http://www.cnil.fr/en-savoir-plus/textes-fondateurs/loi78-17/

[22] Une vingtaine d'auditions et d'interviews ont été organisées au CNRS (2011-2013).

principe de finalité[23] quand les hypothèses ne sont pas élaborées *a priori*, le principe de proportionnalité[24] si les données nécessaires ne sont pas connues avant leur utilisation, ni même le principe de non conservation, car on ne peut plus détruire les données utilisées à la fin d'une recherche à cause de l'accès ouvert et de la réutilisation.

Dans bien des cas l'utilisation de données concernant des personnes est soumise à des contraintes. Ainsi la vision par ordinateur, qui a pour objet de reconnaître automatiquement des scènes visuelles, est soumise au droit à l'image. Dès lors qu'elle se rapporte à une personne identifiée ou identifiable, l'image d'une personne est une donnée à caractère personnel. Le traitement informatique de cette donnée (numérisation, diffusion à partir d'un site web, etc.) doit donc s'effectuer dans le respect de la loi "Informatique et libertés". La Commission nationale de l'informatique et des libertés (CNIL) donne son accord à des fins de recherche pour une recherche sur la reconnaissance de visages à la condition que ces données ne soient pas conservées au-delà du terme du projet, sauf demande de prolongation. Cela entraîne donc le paradoxe d'interdire d'expérimenter d'autres systèmes sur les mêmes données pour en comparer les performances, ce qui pourtant serait la démarche scientifique normale. Il paraît incohérent de limiter ainsi l'utilisation de ces images, alors que le but est de permettre le développement d'algorithmes les plus efficaces possibles en mode opérationnel.

Il est devenu difficile d'appliquer dans tous les cas les principes de base du traitement des données personnelles, tels qu'informer les personnes sur le devenir et l'utilisation des données, ou obtenir leur consentement. Il peut arriver que la démarche du chercheur impose d'obtenir des informations à l'insu de la personne objet de son enquête. Il serait alors nécessaire de prévoir des principes à respecter s'il n'y a pas consentement, comme l'engagement à informer *a posteriori* cette personne[25]. De même la question du consentement se pose quand les recherches portent sur les informations issues de la fouille de données sur des réseaux sociaux. Ces données, publiquement disponibles, sont considérées par la CNIL comme des données personnelles. Signalons enfin que le futur règlement européen sur la protection des personnes physiques à l'égard du traitement des données à caractère personnel[26] *prévoit que des dérogations à l'exigence de consentement en matière de recherche pourront être permises dans les cas où celle-ci sert un intérêt public majeur et ne pourrait être menée à bien d'une autre façon*[27]. Quant au projet de loi sur le numérique, les premières consultations ont mis dans les priorités la définition d'un cadre légal à la fouille de données (*data mining*)[28].

---

[23] Seules doivent être enregistrées les informations pertinentes et nécessaires pour leur finalité.
[24] On doit se limiter à ce qui est nécessaire à la concrétisation des objectifs en rapport avec la finalité poursuivie.
[25] Ceci se pratique déjà dans le cadre de recherche en psychologie par exemple et est prévu dans le cadre de la loi encadrant les recherches sur la personne humaine lorsque «les exigences méthodologiques de la recherche ne sont pas compatibles avec le recueil du consentement et l'information individuelle de la personne». Cette loi promulguée en 2012 est toujours en attente de décrets d'application.
[26] Résolution législative du Parlement européen du 12 mars 2014 sur la proposition de règlement du Parlement européen et du Conseil relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (règlement général sur la protection des données) (COM(2012)0011 – C7-0025/2012 – 2012/0011(COD))
[27] http://ec.europa.eu/justice/data-protection/
[28] Voir le rapport «Study on the legal framework of text and data mining» (mars 2014) - http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf

E. L'ouverture des données de la recherche et le mouvement des communs scientifiques

Depuis quelques années, la naissance d'une science des données et l'explosion des données massives ont conduit à prendre conscience des verrous juridiques et technologiques qui font obstacle à la libre circulation des données, même quand les scientifiques y aspirent. En effet, soit les grandes bases de données sont soumises à des droits d'accès, soit les données sont disponibles sous des formats fermés, nécessitant des logiciels propriétaires. Des initiatives ont alors incité certaines communautés scientifiques à s'organiser pour affirmer des principes nouveaux d'ouverture et de disponibilité des données. C'est ainsi qu'en 2005 une communauté de chercheurs, conscients des résistances rencontrées lors de la mise en place des politiques d'ouverture des données, avait lancé une initiative globale visant à créer concrètement des Communs Scientifiques (Science Commons) avec des outils et des méthodes (plateformes d'accès, contrats d'auteur type,...) pour accélérer la circulation des résultats et permettre la réutilisation des données sur lesquels ils sont fondés[29]. D'une façon générale aujourd'hui, les données pour être partagées font l'objet d'un questionnaire minimum sur l'identification du chercheur et sur la finalité des données.

Ces plateformes de recherche partageables[30] facilitent en effet le développement de nouveaux services comme *la réutilisation des données de la* recherche grâce à des politiques et des outils qui aident les individus et les organisations à rendre leur production accessible ; l'accès immédiat à des outils (calculs en ligne) grâce à des contrats standards permettant de dupliquer, vérifier, étendre les recherches et aussi d'en faciliter l'examen scientifique par les pairs ; l'intégration de sources fragmentées d'information à travers un langage commun, standardisé et traduisible dans la machine.

Le mouvement général d'ouverture et de partage des données est facilité par des politiques d'archivage ouvert développées au sein des institutions scientifiques (ArXiv, 1991). En France, HAL (Hyper Article en Ligne), créé en 2000, est fondé sur «le modèle de la communication directe entre chercheurs»[31] de leurs articles pré-publiés : son pilotage et ses missions, en cours de refonte, sont encore à définir vis-à-vis de l'archivage des données scientifiques ; il s'agira en particulier d'intégrer directement sur la plateforme le lien avec le délai d'embargo et la licence ouverte, dont le choix dépend uniquement du chercheur lors du dépôt.

Un exemple réussi de mise en réseau ouvert initié par des chercheurs est une initiative datant de 2013 issue du domaine de la biologie humaine : la «Global alliance for genomics and health»[32]. Il s'agit d'un mouvement unifié de 285[33] institutions membres de

[29] Aujourdhui le projet, initié par John Wilbanks est développé par Sage. Voir aussi : http://sciencecommons.org/about/

[30] Ce fut le fondement du projet Science Commons voir : D. Bourcier, Web, «Science et Communication : l'exemple de Science Commons», *Revue Hermès* 57, 2010, pp. 53-160 - http://www.cairn.info/zen.php?ID_ARTICLE=HERM_057_0153

[31] Voir Rapport Serge Bauin, *L'Open accès à moyen terme : une feuille de route pour HAL,* DIST, CNRS, septembre 2014 - http://corist-shs.cnrs.fr/sites/default/files/billets/cnrs_dist_rapport_bauin_sur_ccsd_et_hal_septembre_2014.pdf

[32] *Nature* 498, 16–17 (05 June 2013) | doi:10.1038/498017a. L'initiative http://genomicsandhealth.org/ a produit un «Framework for Responsible Sharing of Genomic and Health-Related Data», qui a au préalable fait l'objet de

30 pays qui ont décidé de contribuer à faciliter le partage des données scientifiques par l'établissement de standards, le partage d'expérience et de bonnes pratiques et l'établissement d'un cadre pour le partage responsable des données. De même une autre initiative internationale initiée et pilotée en France dans le domaine de la biologie (BRIF)[34] promeut le partage par une meilleure reconnaissance des ressources partagées et de leurs auteurs, Des disciplines comme les sciences de la Terre et de l'Espace insistent sur d'autres impératifs qui requièrent aussi l'archivage et la diffusion libre de données : l'observation pérenne des phénomènes naturels met en jeu des processus dont les constantes de temps peuvent être grandes par rapport à la vie humaine ; les données issues de ces mesures sont par essence non reproductibles et sont à la base de nos connaissances sur le monde qui nous entoure, de ses évolutions et des risques qui pèsent sur nos sociétés. La nature étant un bien commun, ceci implique vis-à-vis du public une obligation pour la communauté des chercheurs concernés et pour les organismes qui les emploient.

D'autres types de plateformes de données se constituent parfois en dehors de l'initiative des chercheurs. C'est le cas dans des disciplines comme la biologie et la médecine, où les éditeurs exigent des chercheurs qu'ils fournissent leur jeu de données pour vérifier la reproductibilité de l'expérience ou du processus faisant l'objet de la publication, afin de contrôler les résultats à publier en les confrontant aux données et de détecter des fraudes ou des erreurs éventuelles pouvant entraîner la rétractation[35]. Une fois accumulées, ces données, si elles ne sont pas dans des bases publiques (cas décrit dans la note 34), mais restent exclusivement entre les mains des éditeurs, risquent de constituer pour les éditeurs un «marché de données» fermé et autonome par rapport aux publications, alors qu'elles n'étaient demandées que pour contrôler les résultats. Pour éviter cette éventualité, il est nécessaire que l'accès aux données relatives à une publication exigées par les éditeurs ne puisse être limité par le copyright imposé par la revue au texte. Elles doivent être fournies avec l'article en fichier associé mais elles doivent rester à la disposition des chercheurs pour la répétition des analyses et la publication dans n'importe quelle revue.

Les politiques de partage impliquent d'informer les chercheurs sur les limites de ce partage. Les données concernées peuvent être indisponibles en tant que données personnelles non anonymisées, ou peuvent être soumises à des régimes particuliers comme celui de la sécurité nationale et du secret professionnel, à des clauses contractuelles restrictives ou à divers intérêts commerciaux. En outre, le chercheur doit rester titulaire de droits sur les données qu'il produit ou qu'il analyse comme sur ses publications, s'il souhaite les partager ou en autoriser la réutilisation. Dans ce cas, il lui est vivement conseillé de

consultation publique et est disponible aussi en français (http://genomicsandhealth.org/about-the-global-alliance/key-documents/framework-responsible-sharing-genomic-and-health-related-data).

[33] Initialement 50 institutions de 8 pays lors de la naissance de l'Alliance.

[34] L'initiative BRIF (Bioresource research impact factor : https://www.bioshare.eu/content/bioresource-impact-factor) a pour objectif de favoriser le partage des bioressources par une meilleure reconnaissance du travail mené pour constituer, maintenir et rendre partageables ces ressources, échantillons biologiques, données et logiciels associés (Mabile et al. *Quantifying the use of bioresources for promoting their sharing in scientific research*. *GigaScience* 2013, 2:7) ; ce groupe a récemment publié en open access une recommandation pour la citation de telles ressources (Bravo E et al. Developing a guideline to standardize the citation of bioresources in journal articles (CoBRA). BMC Med. 2015;13(1):266) désormais référencée dans un site de recommandations internationales (http://www.equator-network.org/reporting-guidelines/cobra/)

[35] La banque de données sur les protéines, Protein Data Bank ou PDB (http://www.rcsb.org/pdb/home/home.do) est un exemple de ressource accessible dans le domaine public dont la consultation est gratuite. Elle rassemble les données sur la structure tridimensionnelle des macromolécules déposées par des biologistes du monde entier. Les éditeurs de journaux scientifiques et de nombreuses sources de financement réclament le dépôt des structures dans la PDB lors de la soumission des publications ou des projets.

mettre ses données protégeables sous une licence libre comme Creative Commons[36] pour en informer les futurs utilisateurs. Les chercheurs doivent être vigilants sur les conséquences de leur choix quand ils cèdent à des tiers leurs droits d'exclusivité[37].

F. La responsabilité du chercheur

Aujourd'hui, malgré les incitations européennes et celles du CNRS, les chercheurs n'ont pas tous les mêmes pratiques ni les mêmes contraintes vis-à-vis de leurs données, comme le souligne l'enquête effectuée au CNRS auprès des directeurs de ses laboratoires[38]. L'embargo de six mois à un an suivant les délais de publication serait dans les sciences humaines et sociales le délai minimum pour que les données primaires soient rendues disponibles. En chimie, sans doute en raison des questions de valorisation industrielle, la communication des résultats et des données ne précède pas l'acceptation de la publication. Par contre les physiciens mettent en général en accès libre sur des archives ouvertes leurs articles, avec éventuellement des données complémentaires, dès leur soumission, voire avant. Pour l'exploitation des données traitées issues des grands instruments de la physique ou de l'astronomie, il y a un délai avant leur mise au service de toute la communauté, fixé par avance (entre un et deux ans), en donnant une préférence pour un temps limité aux chercheurs ayant construit un instrument sur le grand équipement.

D'une façon générale, le chercheur public est incité à poursuivre un idéal de partage et d'échange entre pairs et à participer à la diffusion des données obtenues sur fonds public, à condition de respecter les exceptions issues d'engagements contractuels. Inversement, les modèles d'accords de consortium impliquant des partenaires publics et privés (notamment dans les Pôles de compétitivité) sont souvent très restrictifs en matière d'ouverture de données : ils devront désormais être négociés en amont par les chercheurs publics de façon à ne pas conduire à une confiscation indue des données non exploitées par les partenaires privés. Une autre préoccupation concerne les modalités de la mise à disposition des bénéfices du partage, pour éviter que les institutions puissantes ou des compagnies privées en tirent un avantage exclusif. On a vu en effet que des équipes financées par des fonds publics ou par l'Union européenne divulguent désormais leurs données alors que les grands groupes privés peuvent les exploiter pour leur propre bénéfice, sans conditions de réciprocité vis à vis des chercheurs publics.

Les chercheurs prennent conscience que l'ouverture des données – mais aussi des logiciels, des ontologies et des métadonnées qui en permettent l'exploitation – implique une nouvelle responsabilité : celle d'être particulièrement soucieux de la qualité des informations et des données qu'ils offrent, ainsi que de la clarté de la documentation qui les accompagne. Pour permettre à d'autres de répliquer ou de réutiliser des données, il est nécessaire de vérifier le caractère intègre et interopérable des données, l'identification de leurs sources, leurs dates de recueil ou de traitement, ainsi que l'examen détaillé des différentes étapes de

[36] www.creativecommons.fr qui est un projet de partage de contenus et une plateforme de licences ouvertes dont les options s'étalent de la licence la plus ouverte (mention de l'attribution) à la plus "marchande".
[37] Le statut des banques de données en Europe, défini comme *sui generis*, n'est pas transposable dans la plupart des autres pays. Les Etats-Unis par exemple ne reconnaissent pas le droit d'auteur sur les banques de données.
[38] Enquête *«Mieux partager l'information scientifique pour mieux partager les connaissances»* - http://www.cnrs.fr/dist/z-outils/documents/Enquête DU - DIST mars 2015.pdf

la constitution de dépôts de données : collecte, classification, standardisation, mise à disposition, réutilisation, conservation, destruction ou archivage. Ainsi l'organisation et la maintenance de données interopérables deviennent des moments fondamentaux pour garantir l'intégrité des données scientifiques à l'heure numérique. Ces nouvelles tâches créent de nouvelles responsabilités *entre les chercheurs*[39]. Il convient d'apprécier au cas par cas les implications de ces politiques face aux dimensions éthiques de la recherche.

En résumé, face à cette dynamique de circulation des données relayée par leurs autorités de tutelle et par leur communauté, les chercheurs doivent prendre conscience de leur responsabilité individuelle, déontologique et éthique, vis à vis de la communauté à laquelle ils appartiennent, avoir connaissance des engagements internationaux des institutions dont ils dépendent, connaître les limites des techniques d'exploitation des masses de données qu'ils utilisent et les difficultés d'interprétation qui en résultent. Il leur appartient aussi de participer à la définition de bonnes pratiques propres à leur discipline dans le domaine du partage des données. Pour ces raisons, les établissements de recherche devront mettre en place des compétences nouvelles qui répondent aux besoins d'information des chercheurs quant à leurs données et créer des comités d'éthique sur les données de recherche par discipline ou établissement.

[39] «Ensuring the integrity, accessibility and stewardship of research data in the digital age», Report of Committee of Science, Engineering and public policy, Washington, The National Academies Press, 2009 - http://www.nap.edu/openbook.php?record_id=12615&page=R1

III. RECOMMANDATIONS

1. Le COMETS rappelle que le CNRS est signataire de la Déclaration de Berlin (2003) sur l'ouverture des données de la science, comme la plupart des grands organismes de recherche au plan national et international. La politique de l'institution engage les chercheurs dans le mouvement mondial de partage de données scientifiques ouvertes. Le COMETS invite tous les acteurs de la recherche des unités du CNRS à s'associer à ce mouvement dans le respect des pratiques propres à chaque discipline. En particulier dans certaines disciplines, des embargos limités à l'exploitation des données par le chercheur peuvent être demandés temporairement ainsi que des garanties minimales concernant l'identification du chercheur ou de l'équipe qui utilisent les données ainsi que la finalité de leur recherche.

2. Le COMETS préconise qu'une réflexion sur la spécificité des données scientifiques et de leur traitement soit menée avec la CNIL et le correspondant Informatique et Libertés du CNRS, ainsi qu'avec ETALAB, les autres organismes de recherches concernés (CEA, INSERM, INRIA, etc.) et la CERNA (Commission de réflexion sur l'éthique de la recherche dans les sciences du numérique d'Allistène). Il suggère la création d'un Comité consultatif d'administration des données de recherche, impliquant diverses disciplines dans cette réflexion.

3. Les chercheurs et les personnels du monde de la recherche doivent être formés aux dimensions éthiques de la gestion des données, en particulier au respect de la vie privée, de la propriété intellectuelle, de la qualité et de l'intégrité des données. Ils doivent être informés de l'état actuel et de l'évolution des règles juridiques concernant le partage responsable de données utilisées.

4. Le COMETS souligne l'importance 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.

5. Le travail de mise à disposition de données utilisables à partir de données brutes doit être reconnu dans l'évaluation et les décisions de promotion des personnels qui s'y impliquent. Pour faciliter cette reconnaissance, le COMETS préconise qu'une rubrique sur ces activités soit ajoutée dans le rapport d'activité et la fiche annuelle d'activité des chercheurs. De même, les auteurs des publications devront citer les chercheurs qui ont contribué aux données et signaler correctement les sources des données qu'ils ont utilisées.

6. Les chercheurs doivent être vigilants quand ils cèdent à des tiers les droits d'exclusivité sur leurs données ou sur des banques de données. Il est nécessaire que l'accès aux données relatives à une publication ne puisse être limité par le copyright imposé par la revue au texte. Elles doivent être fournies avec l'article en fichier associé mais elles

doivent rester à la disposition des chercheurs pour la répétition des analyses et la publication dans n'importe quelle revue.

7. Le COMETS préconise que les archives ouvertes HAL soient privilégiées pour le dépôt des données sur lesquelles s'appuient les publications des résultats de la recherche, s'il n'existe pas déjà de base internationale ouverte dédiée à ces données. Le chercheur doit pouvoir choisir, par des licences ouvertes telles que les Creative Commons, les conditions de leur réutilisation.

8. Le COMETS propose que le CNRS favorise l'acquisition de compétences ou de services nouveaux qui répondent aux besoins d'information des chercheurs quant à leurs données et crée des comités d'éthique sur les données de recherche par discipline avec d'autres organismes.

9. Le COMETS suggère que le CNRS incite ses chercheurs à participer aux instances internationales de normalisation pour traiter des métadonnées. Il encourage fortement l'utilisation des identifiants uniques et persistants afin de permettre le traçage et l'interopérabilité des données.

7 mai 2015.