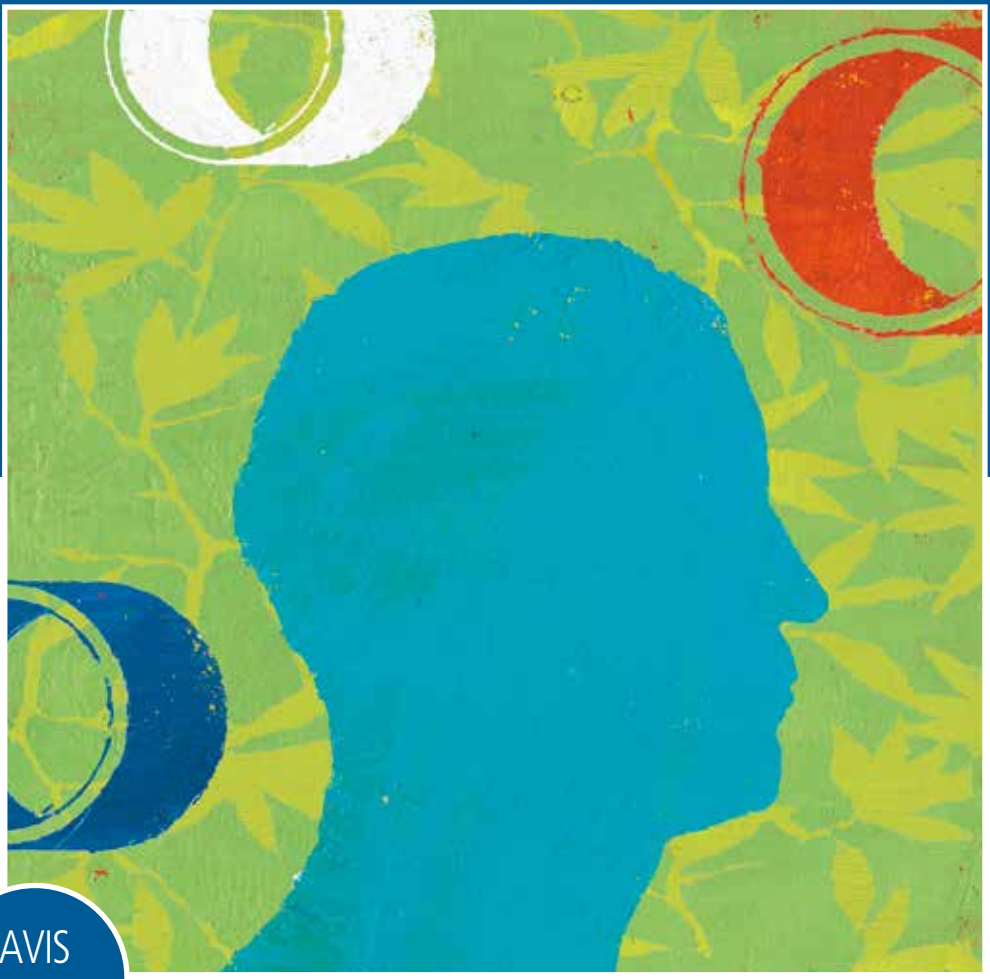
Avis 8 sur les enjeux éthiques et déontologiques du partage et de la gestion des données issues de la recherche

. Comité Consultatif Commun d'Éthique Pour La Recherche Agronomique

Comité consultatif commun
d'éthique pour la recherche agronomique



AVIS
8

SUR les enjeux éthiques et déontologiques du partage et de la gestion des données issues de la recherche

INRA
SCIENCE & IMPACT

cirad

Comité consultatif commun
d'éthique pour la recherche agronomique

AVIS 8

SUR les enjeux éthiques et déontologiques du partage et de la gestion des données issues de la recherche

Sommaire

Résumé de l'avis

Le partage des données de la recherche désigne la mise à disposition sur le réseau informatique mondial (Internet), d'enregistrements factuels sous forme numérique (selon le contexte : résultats d'expériences, observations de terrain, résultats d'enquêtes, images, etc.), collectés pour être analysés dans le cadre d'un travail de recherche.
Cette initiative a pour objectif premier d'accroître la transparence de la démarche scientifique, un mouvement dénommé « science ouverte ». Cette évolution s'intègre dans un dessein plus vaste, dont l'ambition est de faciliter l'accès des citoyens aux diverses informations produites par les gouvernements et les administrations publiques, car ces informations constituent, de fait, un « bien commun ».

Le partage des données issues de la recherche a aussi pour but déclaré d'en faciliter la réutilisation. Ce sujet a suscité de nombreuses analyses, notamment juridiques à propos de la propriété intellectuelle (chercheur ou institution) ou de la protection des données personnelles. Il a aussi donné lieu à l'évaluation de potentielles retombées économiques liées à l'accroissement de l'efficacité de la recherche (point de vue de certains bailleurs) et, au sein des institutions de recherche, à des interrogations sur la « révolution » méthodologique qui pourrait en découler (recherche pilotée par les données, plutôt que par des hypothèses). En comparaison, les implications déontologiques et éthiques liées à la gestion et au partage des données issues de la recherche ont été assez peu explorées.
Le partage des données suppose un travail rigoureux qui commence dès la conception d'un projet de recherche et peut se prolonger bien au-delà de son achèvement. En effet, la mise en ligne de données doit être précédée d'une délicate étape de préparation et d'organisation pour les rendre non seulement accessibles, mais aussi intelligibles en dehors de leur contexte d'origine. Produire des données de grande qualité représente une exigence déontologique majeure pour chaque chercheur. Cependant, le partage des données implique aussi de pouvoir fournir la preuve de leur fiabilité, et par conséquent de développer un contrôle qualité exhaustif. Enfin, la réutilisation des données ne sera possible que si elles sont « portables », « inter-compatibles » et « interopérables », ce qui implique un considérable effort de standardisation qui ne pourra être mené à bien qu'en s'appuyant sur des collaborations internationales.

La recherche a pour objectif premier de produire des données à l'origine de nouvelles connaissances. Cependant, toutes les données issues de la recherche ne peuvent être partagées sans discernement. Compte tenu de l'extrême diversité des activités de l'Inra et du Cirad, de la complexité des choix qui en résulteront (de la mise en ligne sans restriction à la conservation hors ligne, en passant par le partage des données selon des conditions très spécifiques), le Comité estime qu'un processus de décision, appliqué au cas par cas, est inévitable.

Le Comité constate que nombre de chercheurs expriment leurs incertitudes sur le statut des données qu'ils produisent, en particulier au regard de cette question du partage et des missions de valorisation de l'Inra et du Cirad. Ils formulent clairement un besoin d'informations et de concertation et attendent de leur institution qu'elle les guide dans ce nouveau domaine, vœu qui paraît d'autant plus légitime que la responsabilité de toutes les données produites appartient, de fait, aux organismes de recherche.

Les recommandations qui suivent sont en cohérence avec celles émises par le Comité d'éthique du CNRS (Les enjeux éthiques du partage des données scientifiques, mai 2015). Elles ont comme objectif de fournir des pistes d'action, dans le prolongement des réflexions et des travaux engagés depuis 2013 dans les deux instituts (Chantier gestion et partage des données pour l'Inra, Patrimoine Numérique pour le Cirad).

Recommandations à l'adresse des directions et des personnels des deux établissements

1• Le Comité d'éthique ne peut que conforter l'Inra et le Cirad dans leur volonté de partager les données issues de leurs recherches. Le Comité d'éthique recommande de définir une politique d'établissement concernant la gestion et le partage des données qui affiche clairement les engagements des organismes et précise les rôles et responsabilités des différents acteurs. La politique d'établissement devra prévoir un important volet d'information et de formation continue des chercheurs à la mise en ligne des données et à leur réutilisation responsable.

2• Il paraît essentiel que la définition de cette politique d'établissement soit étroitement concertée avec les autres organismes français de recherche publique (CNRS, Inserm, IRD, Ifremer…) et les universités, afin d'aboutir à une réelle mutualisation d'expertise.

3• Il convient de garantir l'harmonisation de la sémantique, la standardisation des formats, des métadonnées, du contrôle qualité, etc. Ce travail devra être mené notamment au sein d'ELIXIR, de *Research Data Alliance ou du Global Open Data for Agriculture and Nutrition* (GODAN).

4• La politique d'établissement doit aussi évaluer précisément et dégager les moyens humains et techniques nécessaires pour en assurer le succès. En particulier, s'imposent la création et la reconnaissance de nouveaux métiers spécialisés dans le contrôle qualité, l'édition, la gestion et la mise en ligne de données (conservateurs de données), ainsi que dans l'exploitation de ces données (fouille de textes et de données, FTD).

5• La qualité des données constitue un aspect essentiel de leur fiabilité. Les chercheurs devront y consacrer une attention particulière, mais les organismes devront renforcer et généraliser les procédures de contrôle de cette qualité, ce qui pourra nécessiter des actions spécifiques de formation.

6• Les activités de production et de partage des données de qualité devront faire l'objet d'une reconnaissance des institutions et être prises en considération dans l'évaluation des chercheurs, des équipes, unités et organismes.

7• Les chercheurs travaillant avec leurs partenaires de pays n'ayant pas une politique de gestion des données suffisamment avancée s'attacheront à assurer les formations pertinentes et à veiller à appliquer les standards de protection des données les plus contraignants.

8• Face à la complexité du sujet, à la diversité des choix possibles et à l'exigence de cohérence à long terme, le Comité d'éthique estime nécessaire de créer une instance spécialisée traitant des questions soulevées par la gestion et le partage des données, instance qui aura un rôle d'arbitrage (au cas par cas) et qui établira progressivement une jurisprudence d'où les principales règles concernant le partage des données pourront être tirées (commission des ressources numériques, Corenum).

Son rôle implique la présence de chercheurs, mais sa composition devra être beaucoup plus large, incluant les « conservateurs de données » et des personnes à même d'apprécier les interactions avec les aspects non scientifiques, notamment éthiques, administratifs, juridiques, financiers et économiques.

Cette instance contribuera à la définition de la politique des établissements, à l'élaboration de règles concernant le partage des données et à leur nécessaire actualisation en fonction des évolutions du domaine. Cette instance évaluera également le caractère sensible des données (données personnelles, données relatives à la sécurité collective, données concernant la santé ou l'environnement…). Elle jouera un rôle de référent en matière de gestion et de partage des données pour les chercheurs et les directions de département.

9• Cette instance évaluera, au cas par cas et en accord avec les groupes de chercheurs ou les chercheurs concernés, si elles peuvent être partagées et, le cas échéant, sous quelles conditions.

10• En cas de demande de partage de données pour un usage commercial, cette instance veillera, en lien avec l'administration de l'institut concerné, à ce que leur utilisation fasse l'objet de conditions d'usage précises, définies dans un contrat garantissant, d'une part, les droits du fournisseur de données (établissement) et, d'autre part, la valorisation du jeu de données (chercheur et collaborateurs).

Introduction

Cette nouvelle question de la gestion et du partage des données traitée par le Comité d'éthique est au cœur d'une actualité qui dépasse le cadre même de la recherche. La dynamique d'acquisition massive des données et de leur partage devenant irréversible, quelle liberté avons-nous face à cette ouverture imposée ?

Les termes de la saisine déclinaient les conséquences et les perspectives qui en découlent pour l'établissement de recherche, comme producteur de données, comme utilisateur de données produites par d'autres, mais aussi en tant que gestionnaire de ce que l'on appelle les métadonnées constituant le cahier des charges de l'observation. Pour le chercheur qui considère parfois que son travail est un produit fini, il s'agira alors de le convaincre que sa recherche est aussi le début d'une autre histoire ! La découverte est, certes, motivation passionnante, le recueil des données étant une dimension essentielle du travail scientifique, mais le fait que des données soient rendues publiques crée de nouvelles contraintes et une nouvelle responsabilité, car ces données seront exploitées par d'autres, ce qui nécessitera de formaliser leurs conditions de recueil et de garantir le cadre dans lequel elles sont prises.

Le Comité a d'ailleurs pris soin de s'entendre sur les mots et d'abord sur ce que représente le terme « données ». Une donnée d'observation a besoin d'être décrite par des métadonnées ; elle a besoin d'être environnée pour éviter d'être sujette à la controverse et pour avoir une valeur scientifique incontestable. Détacher la donnée du contexte de sa production pour la valoriser deviendrait, de ce fait, problématique. Une donnée ne devient donnée qu'après un choix, une identification, un enregistrement, et c'est toujours un acte de création et de responsabilité qui existe indépendamment de la finalité qui a conduit à la rechercher ou à la collecter.

La donnée est le résultat d'une activité humaine dans un cadre qui associe des compétences scientifiques et des savoir-faire techniques. Cette combinaison d'investissements intellectuels et matériels, parfois très lourds, mettant en jeu divers métiers, actuels et nouveaux, leur confère non seulement une valeur propre, mais, dans certains cas, une valeur patrimoniale qui dépasse les seuls intérêts de l'établissement de recherche, a rappelé le Comité.

Si, d'emblée, tout le monde est pour l'*open data,* et le Comité a, lui aussi approuvé le principe de partage des données, ce sujet est cependant complexe, parce que des protections de droits individuels doivent être prises en compte, parce que partager toutes les données est certainement une utopie, parce que le partage des données pose également des problèmes d'appréciation, notamment des risques liés à l'exploitation par d'autres de ces données. Le principe de l'ouverture des données appelle, en effet, une jurisprudence d'appréciation, certainement pas aisée à définir, et que tous les chercheurs ne peuvent pas maîtriser.

Le Comité d'éthique a donc estimé qu'il serait judicieux pour l'Inra et le Cirad de créer un comité consultable sur ces sujets de partage et facilitant la construction d'une culture sur l'application de cette règle générale, pour qu'un chercheur, une équipe de recherche, puissent savoir comment se positionner, à quel moment et sous quelles conditions publier les données.
Dans l'accompagnement de ce travail créatif et important, le Comité a aussi recommandé aux directions des deux établissements de définir - et de partager avec l'ensemble des personnels - une politique d'établissement dans ce domaine, en concertation avec les organismes français de recherche publique et les universités. Accompagner, c'est former, partager des critères de qualité, reconnaître la valeur créative de ces activités, apprendre cette culture de partage qui se développe.

L'acquisition d'un déluge de données, l'application d'algorithmes qui produiront de nouvelles informations, voire de nouvelles connaissances, changent-elles la démarche déductive du chercheur ? Sans répondre directement à cette question épistémologique, le Comité constate que l'Inra et le Cirad collectent et analysent des données dans des champs multiples qui constituent une réelle richesse en agriculture, alimentation et environnement et un enjeu économique, mais aussi éthique, et qui ouvrent la possibilité d'identifier de nouvelles questions de recherche pour comprendre et pas seulement mieux décrire ; cette dualité est essentielle à préserver.

Louis Schweitzer
Président du Comité d'éthique

Avis sur les enjeux éthiques et déontologiques de la gestion et du partage des données issues de la recherche

L'ouverture des données de la recherche (résultats d'expériences, observations de terrain, résultats d'enquêtes, images, etc.) met à disposition sur Internet un déluge d'informations, susceptibles d'être analysées et traitées dans le cadre d'un travail de recherche, dans le cadre d'un mouvement dénommé « science ouverte » (*open science*).

Le partage des données issues de la recherche a aussi pour but d'en faciliter la réutilisation. Les implications déontologiques et éthiques liées au partage des données de la recherche ont été jusqu'à présent peu traitées.

C'est l'objectif de cet avis du Comité d'éthique que d'apporter une contribution à cette réflexion au cœur de l'actualité.

1 ■ Présentation de la saisine

L'ouverture des données de la recherche désigne la mise à disposition sur le réseau informatique mondial (Internet), d'enregistrements factuels sous forme numérique (selon le contexte : résultats d'expériences, observations de terrain, résultats d'enquêtes, images, etc.), collectés pour être analysés et traités dans le cadre d'un travail de recherche. Cette initiative s'intègre dans un ensemble d'actions destinées à accroître la transparence de la démarche scientifique, mouvement dénommé « science ouverte » (*open science*).

Le partage des données issues de la recherche a aussi pour but déclaré d'en faciliter la réutilisation. Ce sujet a suscité de nombreuses analyses, notamment juridiques - à propos de la propriété intellectuelle (chercheur ou institution) ou de la protection des données personnelles, au sujet de l'accroissement de l'efficacité d'exploitation qu'on pourrait en attendre (point de vue de certains bailleurs), de la « révolution » méthodologique qui pourrait en résulter (recherche pilotée par les données, plutôt que par des hypothèses) ou des potentielles retombées économiques[1].

En comparaison, les implications déontologiques et éthiques liées au partage des données de la recherche ont été peu traitées[2]. Pourtant, la première interrogation porte sur le partage des données lui-même. Doit-on partager toutes les données ? En cas de partage, celui-ci doit-il être libre de droit ou, au contraire, soumis à certaines conditions ?

La seconde vague de questions concerne les personnes qui conduisent la recherche et celles qui y participent. Les données sont en effet le tout premier fruit du travail des chercheurs. Les en dessaisir sans aucune contrepartie pourrait porter tort à leurs institutions et aux chercheurs eux-mêmes, dont la carrière s'appuie entièrement sur cette production. Ce risque touche également les collaborateurs, d'autant plus qu'ils peuvent être dépourvus des capacités de les exploiter par eux-mêmes, comme cela peut être le cas dans certains pays. Ce dernier point concerne particulièrement le Cirad qui cherche les moyens d'assurer un caractère équilibré et équitable à toutes ses collaborations.

La troisième série d'interrogations implique les institutions de recherche. Quelle organisation devront-elles adopter pour assurer la mutualisation de la production des données, au même titre que celle des méthodes et des outils de la recherche ? Quels moyens mettre en place pour analyser les données produites et décider de leur mise en ligne ? Quelles assurances prendre afin que les données acquièrent une valeur intrinsèque, en tant que produit de la recherche, évaluable, référençable et citable ?

Enfin, comment assurer un usage pertinent et intègre des données produites par d'autres ? Comment documenter, tracer l'usage de ces données ? Quels moyens mettre en place pour valoriser leur réutilisation ? Il existe donc de fortes tensions entre les exigences de la « science ouverte » et les intérêts des diverses parties prenantes. C'est pourquoi, les directions des deux établissements s'interrogent sur la démarche éthique et déontologique à mettre en place pour « assumer cette nouvelle forme indicible de pouvoir » et ont saisi le Comité d'éthique Inra-Cirad de ces questions (Annexe 1).

2 ■ Éléments de contexte

2•1 Un mouvement de partage des données, accéléré par le développement des outils numériques, dans une période de crise et de mutation de l'édition scientifique

Depuis une cinquantaine d'années, différentes initiatives ont tenté de faciliter l'accès des citoyens (et contribuables) aux diverses informations produites par leur gouvernement et leurs administrations. Certaines de ces informations sont colligées dans des bases de données gigantesques ou mégadonnées (*big data*) dont l'exploitation devrait se révéler extrêmement fructueuse. Cet effort de transparence démocratique a été lancé aux États-Unis par le *Freedom of Information Act*[3] (1966) et, en France, par la loi sur l'amélioration des relations avec le public (dite loi CADA, 1978)[4]. Il est intéressant de noter que la

1 Leonelli S. (2013). Why the current insistence on open access to scientific data? Big data, knowledge production, and the political economy of contemporary biology. Bulletin of Science Technology & Society, 33 : 6-11.

2 L'avis du Comité d'éthique du CNRS (Les enjeux éthiques du partage des données scientifiques ; 7 mai 2015) formule 9 recommandations. http://www.cnrs.fr/comets/IMG/pdf/2015-05_avis-comets-partage-donnees-scientifiques-2.pdf

3 https://www.imls.gov/about-us/agency-reports/freedom-information-act-foia

4 http://www.cada.fr/l-acces-aux-documents-administratifs,1.html

directive européenne 2003/98/CE[5] (révisée en 2013), qui définit les règles de réutilisation des documents issus du secteur public dans les États membres, exclut les universités et les instituts de recherche de son champ d'application.

Un mouvement comparable existe pourtant dans le domaine scientifique, mais il doit s'intégrer à une longue tradition qui codifie, de fait, les échanges entre chercheurs. L'habitude consacrée consiste à rédiger un compte-rendu du travail de recherche qui est alors soumis à une évaluation, sous couvert d'anonymat, par au moins deux lecteurs (les « pairs ») œuvrant aussi dans le domaine. En pratique, le projet d'article est adressé à un journal spécialisé qui choisit les « pairs » (revue à comité de lecture) et gère toutes les étapes de la vie du manuscrit, jusqu'à sa publication. Cette procédure comporte plusieurs inconvénients. Le premier concerne le format du manuscrit, nécessairement limité par les contraintes de la revue. Cette compacité peut d'ailleurs nuire à son évaluation. Le second concerne la diffusion de l'article, tributaire de celle du journal scientifique qui n'est accessible que sur abonnement (*club goods*). Cet aspect est fondamental puisque la reconnaissance du chercheur, donc sa carrière, en dépendent.

La publication en « accès libre » (*open access*) est apparue en réaction aux restrictions de format et, surtout, au coût de plus en plus élevé de l'édition scientifique classique qui limite l'accessibilité des articles (*Budapest open access initiative*, 2002[6]. L'accès libre aux revues à comité de lecture implique leur « mise à disposition gratuite sur l'internet public, permettant à tout un chacun de lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces articles, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale, sans barrière financière, légale ou technique autre que celles indissociables de l'accès et de l'utilisation d'internet ». Cette action collective a créé un nouveau domaine dans lequel l'information scientifique se trouve librement accessible à l'échelle mondiale. Pour les utilisateurs, cette source de connaissance n'est pas exclusive et apparaît libre de rivalité, caractères qui correspondent aux définitions des biens publics (*common*) proposées par Vincent Ostrom et Elinor Ostrom en 1977 (voir Hess et Ostrom, 2007)[7]. Il serait légitime que le fruit des recherches conduites dans les universités et les instituts de recherche publics, le plus souvent avec des financements publics, puisse profiter au plus grand nombre. Ce serait également licite, puisque le Code civil prévoit l'existence d'objets « qui n'appartiennent à personne et dont l'usage est commun à tous » (Art. 714). Le programme cadre de l'Union européenne (Horizon 2020)[8] a d'ailleurs érigé en obligation pour tous les bénéficiaires de ses contrats de recherche la communication complète des résultats : les publications doivent être déposées dans un « entrepôt », puis être mises en libre accès, soit par la voie « verte » (archive institutionnelle ouverte), soit par la voie « dorée » (journaux en ligne). Aux États-Unis, une règle comparable sera appliquée par les principales agences fédérales de recherche[9] à la fin de l'année 2015.

Les choses sont allées encore plus loin dans le domaine de la génétique, particulièrement sous la pression du projet de séquençage du génome humain. Les principaux acteurs du domaine en Europe, aux États-Unis et au Japon se sont accordés sur la nécessité de rendre librement accessibles, dès que possible, sur un site internet public (par exemple, *GenBank*), sous la coordination de l'organisation internationale génome humain (HUGO), toutes les séquences annotées[10] produites par les différents centres financés pour ce projet (principes des Bermudes[11] ; Bermuda I, 1996). Cette initiative est basée sur la conviction, largement partagée, que le génome humain appartient à tous et, en conséquence, sur la volonté d'éviter que des laboratoires ou entreprises ne s'approprient des droits exclusifs (*enclosures*) sur certains gènes ou séquences d'ADN (Bermuda II)[12]. On peut considérer que les principes des Bermudes s'apparentent encore à ceux de l'accès libre, dans la mesure où une séquence d'ADN constitue à la fois le résultat d'une recherche (*research output*) et une « donnée » (séquence primaire ; *research input*). La différence - très importante - est que, dans le premier cas, le produit de la recherche est rendu public avant la publication de l'article correspondant. Le choix éditorial réalisé pour le projet sur le génome humain présente donc une double face : il constitue à la fois la publication des résultats d'une recherche et celle de nouvelles données utilisables pour poursuivre cette recherche. Un certain nombre de revues à comité de lecture ont adopté une position intermédiaire qui consiste à publier les résultats de la recherche, sous condition de la

5 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:FR:PDF

6 http://www.budapestopenaccessinitiative.org

7 Hess C., Ostrom E. (2007). Introduction : an overview of the knowledge commons. In Understanding knowledge as a commons. From theory to practice. Hess C. and Ostrom E. eds, Cambridge: MIT Press, p. 3-27.

8 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

9 Kaiser J. (2015). U.S. agencies fall in line on public access. Science, 348 :167.

10 L'annotation du génome consiste à attacher diverses informations biologiques à une séquence donnée.

11 http://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#2

12 Marshall E. (2001). Bermuda rules: community spirit, with teeth. Science, 291:1192.

communication des principales données (non publiées), ce que les technologies de l'information et de la communication (TIC) permettent aisément. Ce choix ne peut que profiter au processus d'évaluation par les pairs, dans la mesure où les informations additionnelles facilitent l'appréciation de la qualité du travail effectué et de la validité de son interprétation. Cependant, il ne faut pas négliger le fait que l'attitude de certains éditeurs représente un réel danger d'appropriation et de détournement des données.

Une étape supplémentaire a été franchie en proposant d'offrir un libre accès aux données (*open data*) géophysiques et environnementales, partant du constat que l'atmosphère, les océans et la biosphère forment un ensemble intégré dont l'étude justifie l'échange complet et ouvert des données scientifiques acquises par les différents pays[13]. Le partage des données est donc un impératif pour de nombreuses disciplines scientifiques. Il devrait en outre conduire à renforcer la qualité des données (contrôle qualité) et à en garantir la reproductibilité (Arzberger *et al.*, 2004)[14]. La nécessité d'ouvrir les données de la recherche financée sur fonds publics a été plus récemment soulignée dans un rapport de l'OCDE (2007)[15] dans le souci « d'augmenter le rendement des investissements publics », mais aussi parce que les données publiques (données générées par des acteurs publics ou privés intervenant dans le cadre d'une mission de service public) sont un bien informationnel commun (Hess et Orstrom)[7]. Les données de la recherche sont définies ici comme des enregistrements factuels (chiffres, textes, images ou sons) qui sont utilisés comme sources principales pour la recherche[15].

Cette tendance se confirme dans l'appel d'offres Horizon 2020 de l'Union européenne qui intègre un plan de gestion des données comportant leur description, ainsi que les dispositions prévues pour les partager (*Guidelines on open access to scientific publications and research data in Horizon 2020*, 2015)[16]. Bien que le partage des données ne soit pas encore un prérequis pour déposer un projet de recherche européen, certaines agences de financement public en ont déjà fait une condition d'éligibilité hors d'Europe. D'un point de vue scientifique, l'idéal serait un accès aux données de recherche gratuit et irrévocable. Cependant, l'OCDE et la commission européenne (H2020) prévoient des restrictions d'accès liées à la sécurité nationale, la protection de la vie privée, etc. La loi dite CADA (Commission d'Accès aux Données Administratives)[17] permet actuellement aux établissements et institutions d'enseignement et de recherche de fixer les conditions dans lesquelles les informations peuvent être réutilisées, mais l'environnement juridique pourrait sensiblement évoluer avec la transposition prochaine de la directive 2013/37/UE relative à la réutilisation des informations du secteur public. Le principal risque qui pèse sur la liberté des données est ici encore celui d'une appropriation. Par exemple, de fortes pressions s'exercent sur les chercheurs et les laboratoires pour « valoriser » au mieux leurs résultats, c'est-à-dire aboutir à leur exploitation économique, activité considérée par de nombreux acteurs de la recherche, y compris par certains instituts publics, comme une source indispensable de financement qui permettrait de compenser la diminution des ressources allouées par les États. La création, en 2007, d'une agence du patrimoine immatériel de l'État (APIE, rattachée au ministère des finances) destinée à aider l'administration publique à relever les défis de « l'économie de l'immatériel » illustre le poids de ces préoccupations.
De nombreuses considérations, souvent contradictoires, peuvent s'appliquer aux données issues de la recherche, laissant présager que leur gestion ne pourra pas faire l'objet d'une règle simple et universelle.

Des raisons scientifiques justifient la mise en ligne des données. La plus évidente d'entre elles est liée au fait que les données, considérées par les chercheurs comme une part très importante du « produit final » de leur travail (*research output*), constituent aussi une matière première (*research input*) sur laquelle peuvent se fonder de nouveaux processus, tels que la réalisation de méta-analyses ou l'initiative de nouvelles recherches.

Or, les publications ne donnent accès qu'à une faible partie des données engendrées par la recherche. Dans de trop nombreux cas, la description des résultats est incomplète ou biaisée, donc difficilement exploitable. À cela s'ajoute le fait que certains résultats ne sont jamais publiés (près de 50%, 9 ans après

13 National Research Council (1995). Committee on geographical and environmental data. Washington DC, National Academy Press.

14 Arzberger P., Schroeder P., Beaulieu A., Bowker G., Casey K., Laaksonen L., Moorman D., Uhlir P., Wouters P. (2004). An international framework to promote access to data. Science, 303: 1777-1778.

15 Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics (2007).Éditions de l'OCDE, Paris : 29 pages.

16 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (version 2.0 du 30 octobre 2015).

17 La loi n° 78-753 du 17 juillet 1978 reconnaît notamment à toute personne le droit d'obtenir communication des documents détenus dans le cadre de sa mission de service public par une administration, quels que soient leur forme et leur support.
http://www.cada.fr/l-acces-aux-documents-administratifs,1.html

la fin d'une recherche biomédicale). Chalmers et Glasziou (2009)[18] ont estimé que toutes ces insuffisances aboutissent à une perte cumulée d'information dans la recherche biomédicale touchant plus de 85% de l'investissement consenti, soit une perte se chiffrant, aux États-Unis, en centaines de milliards de dollars par an.

Les données comme produit matériel d'une activité humaine

L'efficience du recyclage (cycle de vie) ou de la réutilisation des données dépend d'une bonne appréhension de leur nature réelle et de leur domaine de validité. Lorsqu'un chercheur pose une question dans le domaine des sciences de la vie, il construit un système expérimental permettant d'y répondre. En réalité, il définit une stratégie pour optimiser la détection d'un signal spécifique du phénomène auquel il s'intéresse. Les modèles d'étude - qu'ils soient cellules, plantes, animaux, même de lignée « pure », ou humains - soumis à une intervention identique ne répondront pas de façon semblable. De même, leurs caractéristiques en l'absence d'intervention (situation contrôle) ne sont pas uniformes. C'est la comparaison de ces deux situations (avec et sans intervention) qui révèle généralement le signal à analyser. Aux variations liées aux objets d'expérience, s'ajoutent celles qui sont introduites par les méthodes de mesure utilisées pour observer le phénomène. Cette cascade de procédures aboutit à des valeurs qui ne sont ni absolument précises, ni parfaitement exactes. Il est possible de réduire en partie l'incertitude, par exemple en augmentant le nombre d'observations (augmentation du nombre d'expériences, du nombre d'objets étudiés dans chaque expérience, du nombre de mesures sur chaque objet…). En conséquence, les « données » issues de la recherche sont constituées d'une agrégation (par exemple, une moyenne) de mesures. Il n'existe donc rien de tel que des données brutes ou, plutôt, les données brutes n'apportent pas d'information. Ce n'est que la transformation des données brutes, puis l'analyse et l'interprétation de ce qui en résulte qui peuvent éventuellement conduire à une information nouvelle, c'est-à-dire à une preuve à l'appui d'une allégation relative à un phénomène particulier, autrement dit, à la connaissance. Cette transformation offre aussi l'avantage de véhiculer l'information sous une forme plus compacte.

Dans ce contexte, les données apparaissent comme le produit matériel d'une activité humaine. Leurs caractéristiques physiques, par exemple leur format, leur mode d'expression ou leur support, deviennent aussi importantes pour leur intelligibilité que leur fonction conceptuelle. Qu'il s'agisse du produit matériel d'une activité humaine explique aussi que leur élaboration soit frappée d'un caractère éminemment local. En effet, la production de données scientifiques valides suppose la conjonction des compétences nécessaires à la construction du système expérimental approprié à leur production, ainsi qu'à la façon de les recueillir, puis de les traiter. Offrir l'accès aux seules données expérimentales serait donc tout à fait insuffisant. Leur nouvelle exploitation dépendra en effet d'un savoir-faire au moins équivalent à celui utilisé au cours de leur production. L'identification de ces savoirs, c'est à dire des conditions précises d'acquisition des données, est un préalable indispensable à leur réutilisation. Ces informations doivent être fournies par des « données sur les données » ou métadonnées, indissociables du jeu de données lui-même. Les métadonnées doivent être suffisamment précises et détaillées pour permettre de bien définir le domaine de validité des données qu'elles accompagnent, c'est-à-dire de dessiner la frontière entre les utilisations qu'on peut en faire et des applications qui pourraient se révéler injustifiées ou abusives.

Le partage des données présuppose un travail rigoureux qui commence dès la conception du projet de recherche

La mise en ligne de données doit être précédée d'une délicate étape de préparation et d'organisation pour les rendre non seulement accessibles, mais aussi intelligibles en dehors de leur site de production. La réutilisation de ces données suppose qu'elles soient « portables », « inter-compatibles » et « interopérables » aussi bien sur le plan technique que sémantique[19], ce qui implique un considérable effort de standardisation. La définition de cette mise en forme dépasse largement le cadre d'une institution puisque l'accord de toutes les parties intéressées est souhaitable. Un exemple, impliquant des collaborations à l'échelle mondiale, est représenté par le *Research Data Alliance* (https://rd-alliance.org), créé en 2012, qui comporte de nombreux

18 Chalmers I., Glasziou P. (2009). Avoidable waste in the production and reporting of research evidence. Lancet, 374 : 86-89.

19 http://www.minervaeurope.org/structure/workinggroups/userneeds/prototipo/progproto/interoperabilita_f.html

groupes de travail couvrant tous les aspects de la préparation, de la conservation et de la réutilisation des données. Une initiative comparable - ELIXIR (https://www.elixir-europe.org) - à laquelle participent plusieurs instituts français, a été lancée en 2014 par le laboratoire européen de biologie moléculaire. Une démarche d'ampleur comparable, mais plus spécifique de l'agriculture et de l'alimentation est entreprise dans le cadre du GODAN (*Global open data for agriculture and nutrition* ; http://www.godan.info).
Il apparaît aujourd'hui évident que la méthode de production et de gestion des données destinées à être partagées doit être définie dès la conception du projet de recherche, au même titre que tous les autres choix méthodologiques. C'est ce qui a conduit la Direction générale de la Recherche à exiger, en réponse à ses appels d'offres, un chapitre consacré au plan de gestion des données.
Il ne faut pas sous-estimer les efforts que ce travail va imposer aux chercheurs. On peut prédire que cette charge supplémentaire sera la principale entrave à une mise en ligne effective des données.
L'organisation retenue par les instituts de recherche jouera donc un rôle décisif dans le succès de ce projet. Cela implique tout à la fois de diffuser les guides de bonnes pratiques validés, de former les chercheurs et de fournir les moyens matériels et humains nécessaires, notamment les « conservateurs de données » dont la formation combine de solides compétences dans les domaines scientifiques d'intérêt et dans les technologies de l'information, seuls à même de guider les équipes dans la réalisation de cette nouvelle tâche.

2•2 Le contexte propre à l'Inra et au Cirad

Le document d'orientation de l'Inra (2010-2020) souligne deux originalités de la recherche agronomique : « elle étudie, au laboratoire comme en conditions réelles, un vaste spectre de phénomènes et de systèmes biologiques, écologiques, techniques ou socio-économiques et fait donc appel à un large socle de disciplines. Pour répondre à ces enjeux globaux, la recherche agronomique nécessite, plus que jamais, le recours à des approches systémiques et est concernée par quatre défis majeurs : l'étude des changements d'échelle et de niveaux d'organisation ; la complexité intrinsèque des systèmes étudiés ; l'appel au renforcement des approches inter- et transdisciplinaires ; l'anticipation des futurs changements scientifiques et technologiques et des demandes qui pourront être adressées à la recherche ».

Ces changements touchent notamment l'acquisition des données dont la diversité et le débit continuent de s'accroître de façon spectaculaire (« déluge de données »), posant ainsi des questions nouvelles en matière de gestion, d'analyse et de partage des données. Le premier risque que fait courir un accroissement si rapide du débit de production des données est bien celui de la sous-utilisation d'une grande partie d'entre elles.

Un audit des infrastructures informatiques, conduit entre 2009 et 2011, a documenté de manière exhaustive l'ensemble des composants informatiques scientifiques de l'Institut. L'Inra dispose ainsi d'un capital d'environ 3 500 composants applicatifs, dont presque 860 bases de données, et d'environ 900 serveurs. Dans cet ensemble, 91 d'entre eux (composants applicatifs, bases de données ou infrastructures informatiques) ont été identifiés comme stratégiques pour soutenir les orientations de recherche de l'établissement, chacun ayant été évalué par le collège de direction et les chefs de département en termes d'enjeux et de stratégie[20]. Par ailleurs, le conseil scientifique de l'Inra a conduit une réflexion prospective relative à la gestion et au partage des données et étendue à l'ensemble des domaines scientifiques de l'Institut[21]. Ce rapport mettait en lumière les enjeux scientifiques et politiques de la science ouverte et des mégadonnées dans de nombreux domaines d'intérêt pour l'Institut, ouvrant à la recherche des perspectives de production de nouvelles connaissances par l'intégration de la diversité et de la complexité des données disponibles, de nouvelles collaborations par la valorisation de ses propres données, tout en posant au chercheur des questions d'ordre technologique, stratégique, éthique et juridique, et se concluait par neuf recommandations.

Suite aux recommandations du conseil scientifique de l'Inra, un chantier « gestion, partage et réutilisation des données » a été mis en place en 2013, commun avec le Cirad, et mobilise environ 80 personnes : scientifiques, informaticiens, documentalistes et juristes. À partir de l'ébauche d'un cadre politique

20 Schéma directeur des systèmes d'information de l'Inra (2012), 52 pages.

21 Gestion et partage des données. Rapport du conseil scientifique de l'Inra, adopté le 24 mai 2012 ; 62 pages.

institutionnel au travers de 11 principes[22] élaborés début 2013, le chantier de mise en œuvre du partage des données a pour objectif de faire l'état de l'art sur le sujet, l'inventaire des pratiques dans les établissements et d'établir des recommandations afin de construire une offre de services pour accompagner les équipes de recherche dans le partage et la réutilisation des données de recherche : stratégie de valorisation selon trois familles de données – « omiques », environnement/expérimentation/observation, sciences humaines et sociales (enquêtes, cohortes, discours et représentations), guide juridique et de la propriété intellectuelle, plan de gestion de données, identifiants numériques de jeux de données, choix d'un entrepôt…

C'est dans ce cadre que le Cirad a ouvert un chantier « Patrimoine numérique scientifique » (PNS) dont l'objectif principal est d'organiser, préserver et valoriser le patrimoine numérique du Cirad passé et futur, avec au cœur de ses préoccupations les données scientifiques propres telles que celles issues de la plateforme « analyse du cycle de vie » (ACV) et qui regroupe l'une des équipes les plus importantes en France dans ce domaine.
Il s'agit dans un premier temps de dresser un inventaire du PNS (données et modèles), de constituer un catalogue (base de données de référence), un annuaire des données scientifiques (intranet), afin d'élaborer une cartographie des familles de données et d'identifier les ressources essentielles et critiques.
Afin de sécuriser les futures données produites par le Cirad et ses partenaires, un modèle de plan de gestion de données est en cours de construction. L'objectif est aussi de les valoriser sous forme de publications originales (articles de données), de ressources pour l'enseignement, etc… La sauvegarde dans des entrepôts de données sécurisés revêt un caractère stratégique, notamment en vue d'une réutilisation ultérieure dans le cadre de projets de recherche. Dans le but d'anticiper les changements à venir, de nouvelles pratiques scientifiques numériques doivent être formalisées et adoptées au sein de l'organisme, avec comme visée une préparation aux évolutions des politiques d'État ou des bailleurs de fonds.[23]

Enfin, en termes de gestion prévisionnelle des emplois et des compétences, les besoins inhérents à la conception de méthodes et d'outils d'analyse des données ou à l'acquisition, l'analyse et le traitement des données et, plus généralement, à la conservation, la gestion et la valorisation des données sont largement exprimés par les départements de recherche, puisqu'ils représentent près de 40 % des besoins identifiés lors d'une étude récente portant sur la gestion prévisionnelle des emplois et des compétences à l'Inra.[24]

3 ■ Enjeux pour l'Inra et le Cirad, producteurs et utilisateurs de données

3•1 Une exceptionnelle variété des jeux de données qui appelle un traitement différencié

Les activités de l'Inra et du Cirad touchent un vaste spectre de phénomènes et de systèmes biologiques, écologiques, techniques ou socio-économiques. À titre d'exemple, les recherches menées à l'Inra portent sur l'agro-écologie et les systèmes agricoles, l'environnement, la biologie végétale et animale, les biotechnologies, l'alimentation et la nutrition animale et humaine, mais aussi sur l'économie et les sciences sociales.
Les préoccupations du Cirad les recoupent largement, tout en étant tournées vers les pays du Sud. Elles concernent en particulier les conséquences du changement climatique sur l'agriculture, mais aussi la santé animale et végétale, les structures agricoles et notamment l'agriculture familiale et le pastoralisme, les filières, la biodiversité et le développement durable et, évidemment, la sécurité alimentaire.
La diversité de ces recherches aboutit à la création de très nombreuses bases de données différentes : données de génétique et de génomique (plantes, arbres, champignons, animaux et micro-organismes…), d'expérimentation et d'observation (écologie, climat, paysage, sol, plante, arbre, animal, biodiversité, système de culture, socio-écologie…), de modélisation (environnement, paysage, eau, plante, forêt, animal, socio-écologie et systèmes géographiques), d'enquêtes et de cohortes (alimentation, santé, consommateurs, industries agro-alimentaires, politiques publiques, exploitations agricoles, commerce…).

22 Les 11 principes sont organisés selon cinq axes : (1) principes « cadres » permettant de répondre à l'enjeu de l'« *open science* » ; (2) principes en lien avec la propriété intellectuelle, la déontologie, les droits d'auteur ; (3) principes relatifs aux bonnes pratiques et standards – les données doivent être publiées en cohérence avec les pratiques nationales ou internationales de la discipline dont elles relèvent ; (4) principes de référencement/suivi/traçabilité/visibilité en interne ; (5) principes en lien avec le financement du coût de la gestion de la donnée scientifique. Voir https://intranet6.Inra.fr/systemes-information/Projets-en-cours/Partage-donnees

23 Le Cirad dispose également d'une base de données des publications réalisées par ses chercheurs ou avec des partenaires, Agritrop (agritrop.Cirad.fr), dont la nouvelle version a été mise en ligne en juin 2015 : 90 000 références, 22 000 documents dont près de la moitié est accessible à tous. Ces documents sont accompagnés entre autres de métadonnées simples, et permettent d'identifier tous les contributeurs du Cirad.

24 Étude communiquée aux directeurs d'unité en septembre 2014.

Concernant les données « omiques » par exemple, on peut identifier 19 types de données que l'on distingue en données dites « brutes » (6 types) obtenues directement en sortie d'un équipement de mesure et en données dites « élaborées » (13 types), issues d'une analyse de la donnée « brute », la frontière entre les deux types de données n'étant pas toujours aisée à tracer[25] (voir annexe 2).

Quelques exemples soulignent aussi le caractère stratégique de certaines de ces données. Ainsi, la gestion des bases de données nationales, hébergées au Centre de traitement de l'information génétique pour les animaux d'élevage, enrichies de génotypes, constitue une ressource précieuse pour la compréhension des mécanismes fondamentaux de l'élaboration des phénotypes.

L'Observatoire du développement rural, plateforme collaborative de données créée et gérée par l'Inra en partenariat avec le Ministère chargé de l'agriculture et l'Agence de services et paiement, rassemble des bases de données se rapportant aux mesures de politiques agricoles et agri-environnementales, aux activités agricoles et, plus généralement, au développement rural. Il peut mettre à la disposition des chercheurs une partie de ces informations, tout en respectant les règles de confidentialité. Une démarche qualité a été mise en place pour garantir aux utilisateurs la qualité de la donnée, en suivant les diverses opérations subies par les données dites « brutes » (d'origine administrative ou autre) depuis leur arrivée jusqu'à leur mise à disposition et leur exploitation.

Le projet *Wheat Initiative*, dans lequel l'Inra joue un rôle moteur, repose sur un réseau international d'experts et a pour perspective de construire un système d'information intégré sur le blé et de fournir à la communauté scientifique un accès facile aux données de la génétique et de la génomique de cette plante, ainsi que des outils bio-informatiques performants. L'objectif du projet vise à une meilleure connaissance de la génétique du blé afin d'accroître la résistance aux maladies et les rendements pour répondre à l'augmentation de la demande alimentaire mondiale. À terme, le projet *Wheat Information System* (*Wheat IS*) proposera un portail unique où les scientifiques du monde entier pourront disposer de données intégrées et consolidées, à même d'accélérer leurs recherches sur le blé.

Le Cirad dispose de données uniques et originales de différentes plantes tropicales (dont le palmier à huile, la banane, le cacao ou le café), avec la plateforme ACV qui revêt une grande importance scientifique et économique à la fois pour les partenaires privés, mais aussi pour les pouvoirs publics (Ministère de l'environnement, à l'origine de l'expérimentation nationale de l'affichage environnemental lancée à la suite du Grenelle de l'environnement).[26] Ceci n'est qu'un exemple des multiples bases de données (hévéa, essais génétiques, enquêtes de terrain…) que le Cirad est en train de recenser.

L'un des 13 départements de recherche de l'Inra développe des travaux en mathématiques et informatique appliquées, notamment concernant la gestion et l'analyse de masses considérables de données hétérogènes : faire « cohabiter » des données hétérogènes à l'aide de méthodes informatiques d'intégration ; développer des méthodes et algorithmes pour extraire automatiquement des connaissances à partir de ces données ; mettre en œuvre les possibilités technologiques nouvelles d'observation et de récolte des données à des échelles de résolution spatiale et temporelle sans précédent. Ces savoirs sont considérés actuellement comme critiques, car rares au regard des besoins des organisations publiques comme privées, et stratégiques, car susceptibles de produire un avantage concurrentiel[27].

Au-delà de la grande diversité des thèmes et des conditions de recueil, les données qui résultent de ces travaux diffèrent profondément par leur portée pour la collectivité et, en conséquence, par leurs possibilités de partage. Par exemple, s'il s'agit de données pertinentes pour la sécurité sanitaire des cheptels ou, plus encore, des populations (risque de zoonose), il importe de les partager au plus vite avec les scientifiques et les gestionnaires du risque pour contribuer à parer la menace. Au contraire, si les résultats de la recherche comportent des données personnelles ou permettent d'y accéder par recoupements (par exemple, campagnes d'évaluation des politiques agricoles), il est impossible de les diffuser en l'état. Dans ce dernier cas, les altérations qui seraient requises pour la protection des personnes risquent fort d'ôter tout intérêt aux éléments mis en ligne. Enfin, des données acquises sur le long terme (par exemple, évolution phénotypique de lignées animales ou végétales, systèmes d'observation et d'expérimentation pour la

25 Lors de l'acquisition des données dites « brutes », il existe déjà une analyse réalisée sur la plateforme transformant le signal lu (fluorescence, migration) en une donnée génomique, mais cette analyse n'implique généralement pas l'utilisateur final de la donnée pour qui la donnée fournie par la plateforme est un point de départ.

26 Bilan au Parlement de l'expérimentation nationale, sept. 2013. http://www.developpement-durable.gouv.fr/Bilan-au-Parlement-de-l.html

27 Catlin T., Scanlan J., Willmott P. (2015). Raising your Digital Quotient. McKinsey Quarterly, June 2015. "*Companies should recognize that, in many instances, digital competency matters more than sector knowledge, at least at the early stages of digital transformation" in*: http://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/raising-your-digital-quotient.

recherche en environnement, plantations comparatives de génétique forestière) représentent une réelle valeur patrimoniale dont l'intérêt dépasse largement le cadre des établissements qui les détiennent. Il apparaît donc que la nature même du jeu de données considéré peut, d'une part, affecter les possibilités de sa mise en ligne et, d'autre part, imposer la définition de conditions d'usage spécifiques.

3•2 Des enjeux liés à la mission de recherche finalisée dans les deux établissements

L'Inra et le Cirad ont pour mission de « faire progresser les connaissances et répondre à des questions scientifiques suscitées par des enjeux de société et leurs prolongements opérationnels. Les recherches conduites visent à répondre à des problèmes issus de la pratique de différents acteurs socio-économiques. Cet objectif, qui impose le plus souvent un détour par de la recherche fondamentale avec une production de connaissances génériques, implique pour l'institution ou le collectif de recherche de construire des interactions fortes à la fois avec des partenaires scientifiques et avec des partenaires socio-économiques ou des utilisateurs des résultats de la recherche. Ces interactions interviennent dans les différentes étapes de définition d'une stratégie de recherche, de production des connaissances, de transfert et d'utilisation pour l'action et l'innovation.[28]» Une part non négligeable du financement de la recherche est issue du partenariat (40% de fonds dits compétitifs pour le Cirad et, entre 15 et 20%, pour l'Inra) et cette part est appelée à progresser pour la recherche publique en général[29].

La place centrale de la recherche finalisée met les chercheurs et leurs institutions en présence de partenaires aux profils divers, et aux rationalités spécifiques. Citons les principaux concernés par la problématique de partage.

• **Les financeurs de la recherche :** les États, Ministères de tutelle, Agences d'évaluation, ou institutions supranationales (Commission européenne, OCDE), qui encourageaient dans un premier temps et qui, aujourd'hui, conditionnent l'octroi de subvention à la publication des données de recherche.

• **Les bailleurs privés** qui n'ont généralement pas intérêt à la publication des données.

• **Les partenaires de la recherche** : dans le cas du Cirad, les pays en développement voient dans l'utilisation des infrastructures de traitement, de stockage et de publication de données en ligne une occasion d'inclure leurs établissements de recherche dans la mouvance mondiale, mais non sans crainte. Ainsi, la mise en ligne de données génétiques concernant leurs ressources biologiques propres peut créer des tensions : dans le cadre de la souveraineté reconnue aux États sur leurs ressources génétiques, le Traité International sur les Ressources Phytogénétiques pour l'Alimentation et l'Agriculture (TIRPAA sous l'égide de la FAO – 2009)[30] a instauré un système multilatéral d'accès à certaines espèces permettant de conserver un accès facilité à ces ressources utiles à la sécurité alimentaire, sans contraintes pour les données issues de ces ressources. D'un autre côté, la convention sur la diversité biologique (1992)[31] et le protocole de Nagoya (2014)[32] mettent en place des mécanismes d'« accès et de partage juste et équitable des avantages », selon un système bilatéral, qui permettent de conditionner éventuellement l'accès aux ressources à une mise à disposition privilégiée des données au profit du pays fournisseur. Cette option, qui peut limiter le partage, répond à la crainte que l'exploitation de ces données par des acteurs industrialisés contribue à accentuer les décalages technologiques existants. Quel que soit le système retenu, multilatéral ou bilatéral, l'enjeu est de conserver le caractère « juste et équitable » de la mise à disposition des ressources et des données.

• **Les utilisateurs à but commercial** de la recherche publique qui sont principalement intéressés par les méthodes de gestion et d'analyse des données (veille scientifique et technologique sur les meilleures pratiques), et par les données précompétitives (exemple : lignées commerciales de semence).

• **Les participants aux recherches** qui souhaitent être reconnus dans leurs actions, être informés. Ils peuvent former un réseau assez considérable comme dans le cas du projet Pl@ntnet (Groupe Biodiversité du Cirad : 20 000 participants répartis dans 25 pays).

• **Les instituts techniques français** avec lesquels les collaborations sont nombreuses.

28 On trouvera dans les documents institutionnels, comme dans les actes du séminaire sur la recherche finalisée (2007) des éléments qui résument les missions et les objectifs d'une recherche agronomique finalisée.

29 Bush L. (2014). « Le marché aux connaissances – Néolibéralisme, enseignement et recherche ». Éd. Quae, 159 p. (voir page 82)

30 ftp://ftp.fao.org/docrep/fao/011/i0510f/i0510f.pdf?bcsi_scan_628cd39dca2568d2=1

31 https://www.cbd.int/doc/legal/cbd-fr.pdf

32 https://www.cbd.int/abs/doc/protocol/nagoya-protocol-fr.pdf

• **Les citoyens** qui sont de plus en plus attentifs aux résultats de la recherche agronomique et de ses implications. Avec le développement des mégadonnées et l'utilisation des données des réseaux sociaux, il devient simple d'utiliser les données personnelles à l'insu des individus, sans un consentement *a priori* [33] . Or, les dérives en la matière ont déjà été portées à la connaissance des citoyens normalement informés et raisonnablement attentifs et éclairés[34].

L'organisation de la recherche par projets distincts et limités dans le temps, induite en partie par les modes de financement, introduit une discontinuité dans la collecte et la conservation des données, notamment pour les thématiques développées dans la durée. D'une part, certains sujets de recherche (plantes, évolution des sols, effet du changement climatique, etc.) nécessitent une collecte et un traitement de données qui dépassent de beaucoup la durée limitée des projets. D'autre part, les données collectées sur une période de temps limitée lors d'un travail de recherche devront demeurer accessibles bien après son achèvement, alors même que toutes les ressources mobilisées pour le projet auront été consommées. Il sera donc nécessaire, au niveau des établissements, de planifier sur le long terme la sauvegarde et l'accessibilité des données pour leur réutilisation ultérieure. Cela pose la question de la durée de vie des données, c'est-à-dire de la nécessité éventuelle de les renouveler ou de les détruire passé un certain délai.

4 ■ Questions éthiques et déontologiques soulevées par la publicité des données

Les éléments de contexte mettent en évidence les injonctions à la fois externes et internes auxquelles sont soumis les chercheurs des deux organismes, injonctions qui se révèlent contradictoires lorsqu'il s'agit à la fois de partager les données et de valoriser celles-ci. Les chercheurs ont la perception aiguë d'un univers où la concurrence s'exacerbe avec le développement de la science ouverte qui affecte le travail dans ses phases d'acquisition, d'inventaire et de transformation des connaissances. Il peut en résulter une fragilisation de la position des chercheurs qui doit être prise en compte à chaque stade du déroulement d'un projet ou d'un programme de recherche.

Chaque étape du cycle de vie des données, de leur production à leur réutilisation, soulève des questions éthiques qui lui sont propres.

4•1 Étape de production des données

4•1-1. Le partage des données implique de pouvoir fournir des preuves de leur validité aux futurs utilisateurs, ainsi que des indications sur l'étendue de leur domaine de validité. Cette démonstration dépend de la mise en place d'un contrôle qualité très strict dès le début de la collecte, afin d'assurer une traçabilité exhaustive des conditions de recueil et des méthodes utilisées. Nombre de chercheurs considèrent déjà cet aspect de leur travail comme une exigence déontologique majeure, mais la généralisation de cette pratique constituerait certainement l'un des bénéfices majeurs, et à assez court terme, du partage des données. Le contrôle qualité des données, qui est en cours de déploiement à l'échelle des établissements, suppose une organisation rigoureuse et des normes définissant un crible d'évaluation minimal. L'édiction de ces normes dépasse le cadre de l'Inra et du Cirad, et devrait être le fruit d'une collaboration entre tous les organismes de recherche à l'échelle du pays ou, mieux, de l'Europe, voire du monde.

4•1-2. Si la plupart des chercheurs paraissent plutôt favorables à ce mouvement, en pratique ils n'acceptent de partager leurs données que s'ils peuvent déterminer à qui, quand, comment et pourquoi les transférer. La principale raison de ces réserves est liée à la crainte d'un pillage de leur travail, aboutissant à sa publication par un concurrent qui pourrait en usurper la paternité. Cette crainte est très légitime et pourrait être dissipée en créant des règles claires pour la publication des données, comportant notamment une période de rétention (embargo) d'une durée à déterminer. Ce délai devrait permettre au chercheur d'achever et de publier son travail. Ce point est capital, dans la mesure où le partage des données ne sera effectif que si les chercheurs s'y rallient. Cela

33 La Loi de santé publique votée en avril 2015, dans son article 47, a prévu d'ouvrir les données de la Sécurité Sociale à la recherche scientifique. Toutefois, l'accès sera fonction de certains critères : intérêt public, qualité du projet de recherche, nature des données requises, sécurité des procédures et qualité du demandeur.

34 Harcourt B.E. (2014). Governing, Exchanging, Securing: Big data and the production of digital knowledge. Columbia Public law Research Paper n° 14-390, 37 p. (May, 23) http://ssrn.com/abstract=2443515

implique de recueillir leur pleine adhésion, donc de les associer d'emblée à la construction de la politique de partage (mode de gouvernance et règles) de l'institution. En préalable, une clarification s'imposera : les chercheurs n'ont pas de droits particuliers sur les données qu'ils produisent. Cependant, il semble bien difficile de mettre celles-ci en ligne sans leur consentement. À ce propos, il convient de rappeler que la publication d'un article requiert l'accord préalable de chacune des personnes qui y ont collaboré.

4•1-3. La prise en compte de l'intérêt des personnes qui ont contribué à l'acquisition des données peut aussi conduire à en limiter, au moins temporairement, l'accès. Par exemple, les données sont le plus souvent produites par les chercheurs les plus jeunes et les rendre publiques sans discernement pourrait nuire à leur carrière. En effet, la production et la publication de données issues de la recherche sont bien loin de recevoir la même reconnaissance que les publications. De même, il faut s'interroger sur l'opportunité de la mise en ligne des contributions venant de citoyens (*crowdsourcing*), des agriculteurs par exemple (*vide supra 3.1*). Il importe tout autant d'assurer la protection des données acquises en collaboration avec les pays du Sud, jusqu'à ce que leurs équipes de recherche soient en mesure d'exploiter par elles-mêmes ce qui constitue de fait une part de leur capital scientifique et économique.

4•1-4. Le sort des données concernant les personnes participant à une recherche est beaucoup plus délicat à traiter. La mise en ligne de données personnelles n'est pas possible dans notre pays (Loi informatique et liberté, 1978)[35] et il devient de plus en plus difficile, voire complètement illusoire, de rendre les données personnelles « anonymes » du fait de leur caractère même ou de la possibilité de croiser des bases de données (Mayer-Schönberger & Cukier, 2013)[36]. De plus, les altérations qu'imposerait la protection des personnes ôteraient probablement toute valeur aux données ainsi transformées.

Cependant, l'Agence européenne du médicament (EMA) offre depuis 2010, au cas par cas et sur justification écrite, l'accès à certains documents des essais cliniques sur lesquels sont fondées les décisions de mise sur le marché. Depuis 2014, cette possibilité est étendue à l'ensemble des données des essais cliniques portant sur des médicaments, qu'ils aient ou non reçu une autorisation de mise sur le marché[37]. La gestion des données et de leur partage sera prochainement déléguée à une institution indépendante (*Wellcome Trust*) et leur consultation et réutilisation s'effectuera sur un serveur dédié, muni des outils appropriés, afin de protéger au mieux les données personnelles ou confidentielles et d'éviter toute dissémination incontrôlée. En France, le centre d'accès sécurisé aux données (CASD) permet aux chercheurs de travailler sur des données individuelles très détaillées dans des conditions de sécurité réputées élevées. Toutes ces procédures soulèvent la question du consentement « éclairé » des sujets concernés, dans la mesure où il est impossible d'anticiper les (ré)-utilisations possibles de leurs données.

4•1-5. Les données à caractère dual pourraient être utilisées de façon mal intentionnée, pour nuire aux plantes, aux animaux ou aux personnes (séquence virale, par exemple). Bien que l'exploitation malveillante de telles informations suppose de solides connaissances, leur mise en ligne doit être bien réfléchie. À cet égard, le rapport « *National and transnational security implications of big data in the life sciences* »[38] élaboré en 2014 conjointement par l'*American Association for the Advancement of Science*, le *Federal Bureau of Investigation* et les Nations unies (*United Nations Interregional Crime and Justice Research Institute*) examine les bénéfices et les risques associés au partage ouvert des données issues des sciences de la vie, et propose un cadre méthodologique pour les évaluer de façon continue en prenant en considération les technologies émergentes et « capacitantes » (*enabling*).
Le cas des données révélant des risques pour la santé des plantes (par exemple, risque d'épiphytie), des animaux (épizootie) ou des humains (zoonose) a déjà été abordé (3.1). De façon analogue, les données indiquant une atteinte à l'environnement (pollution) devraient immédiatement être partagées pour que des mesures appropriées puissent être rapidement adoptées.

35 Loi 78-17 du 6 janvier 1978 modifiée. http://www.cnil.fr/documentation/textes-fondateurs/loi78-17/

36 Mayer-Schönberger, V., Cukier K. (2013). A revolution that will transform how we live, work, and think. New York, Houghton Mifflin Harcourt, 256 p.

37 Bonini S., Eichler H-G., Wathion N., Rasi G. (2014). Transparency and the European Medicines Agency – Sharing of clinical trial data. N. Engl. J. Med. 371:2452-2455.

38 http://www.aaas.org/report/national-and-transnational-security-implications-big-data-life-sciences

4•1-6. La gestion et le partage de données qui peuvent s'apparenter à des « biens communs informationnels » ou qui risqueraient de mettre en péril une filière économique confèrent à l'Inra et au Cirad un rôle et une responsabilité spécifiques. L'impact du partage des données doit être évalué très en amont et tenir compte du contexte socio-économique et éventuellement géostratégique. Cela implique de croiser de multiples aspects, au-delà des seules considérations scientifiques.

4•2 Étape de traitement des données et de leur mise en ligne

La question du type de données à mettre en ligne se pose aussi. Il est des situations dans lesquelles le nombre de données saisies par unité de temps atteint des valeurs si vertigineuses qu'il est tout simplement impossible de les colliger toutes, *a fortiori* de les mettre en ligne. Le nombre de désintégrations observées dans chaque expérience du CERN en représente une illustration saisissante. Dans ce cas, c'est seulement une reconstitution partielle qui est accessible.
La mise en ligne de données incomplètes, notamment après exclusion des données personnelles, peut leur retirer tout intérêt. L'exhaustivité des métadonnées, sans lesquelles les données sont inexploitables, soulève la même difficulté. On voit donc apparaître des conflits entre le « souhaitable » et le « possible », en gardant bien à l'esprit que le « conditionnement » même des données mises en ligne peut affecter, voire biaiser leur réutilisation.
La normalisation des présentations, indispensable à la mise en ligne de données homogènes et exploitables, peut aussi provoquer un effet de censure en excluant certains contenus ou certains formats. Une telle contrainte pourrait limiter la liberté du chercheur.
Le choix du site de mise en ligne est très vaste. À cette étape aussi, le risque d'accaparement existe. Il dépend largement de l'attitude qu'adopteront les éditeurs de revues en ligne (voie « dorée »). Il est évident qu'un réel danger de détournement des données existe, si l'éditeur exige la cession des données en échange de leur publication. Il paraît donc préférable de s'adresser en priorité aux entrepôts universitaires ou aux instituts publics de recherche.

4•3 Réutilisation de données mises en ligne

La recherche de données mises en ligne dépendra de logiciels spécialisés et la possibilité existe que cette étape soit orientée dans un sens particulier par les outils disponibles. En effet, la plupart de ces logiciels seront très probablement « propriétaires », gardant confidentiels leurs algorithmes de fonctionnement, créant ainsi une sorte de dépendance. Dans ces conditions, il pourrait se révéler difficile de garantir un accès non biaisé aux données avec des outils utilisant des algorithmes dont on ignore les modèles et les hypothèses. Il est donc important que l'ouverture des codes composant ces outils, quelle que soit leur origine, soit en phase avec l'ouverture des données.

Les données sont le résultat d'une activité humaine qui dépend d'un cadre associant les compétences scientifiques nécessaires à la conception de la recherche et les savoir-faire techniques indispensables à sa conduite. Cette combinaison d'investissements intellectuels et matériels, parfois très lourds, leur confère une valeur propre. Cela est flagrant lorsque les données, concernant des cultivars ou l'amélioration d'espèces animales, ont été acquises sur de très longues périodes. Dans ces cas, elles acquièrent une réelle valeur patrimoniale, susceptible de dépasser les seuls intérêts de l'institut de recherche. Leur mise en ligne ne peut que résulter d'une décision mûrement pesée.

L'exploitation des données publiques dans un but commercial soulève une fois encore la question du risque de détournement ou d'accaparement des résultats de la recherche par une tierce partie. Ainsi des partenaires industriels, parfois n'ayant pas pris part aux projets, peuvent solliciter l'accès aux données obtenues sur fonds publics. Aujourd'hui, il n'existe pas de procédure écrite et les délais de réponse sont souvent très courts, ce qui met les chercheurs dans des situations peu confortables, partagés entre la préservation du patrimoine numérique de leur institut et la nécessité de ménager la relation avec le partenaire. Des organismes mettant à disposition des bases de données (par exemple, *ecoinvent*)[39] peuvent également solliciter un transfert de données vers leur propre base, dont les services sont monnayés. Les chercheurs ont alors l'impression de devoir faire face à un marché obscur, où la donnée devient une marchandise.

39 https://www.ecoinvent.org/database/database.html

Enfin, il convient de s'assurer que la réutilisation des données sera faite de façon responsable, c'est-à-dire en respectant leur domaine de validité et en utilisant une méthodologie adaptée. Leur réutilisation injustifiée ou abusive, ou utilisant des méthodologies inappropriées pourrait conduire à des polémiques sortant du strict cadre scientifique, susceptibles de jeter le discrédit sur le travail réalisé, voire sur une discipline entière.
Cette courte série d'exemples illustre la multiplicité des questions que soulève la publication de jeux de données aussi variés. Nombre de chercheurs expriment déjà leurs incertitudes sur le statut des données qu'ils produisent. Ils formulent clairement un besoin de conseils et de concertation, et attendent de leur institution qu'elle les éclaire sur les enjeux de ce nouveau domaine et sur les choix stratégiques à faire. Ce vœu paraît d'autant plus légitime que la responsabilité de toutes les données produites appartient de fait aux instituts de recherche.

Cette fonction d'information et de concertation doit donc être prévue et intégrée d'emblée à la politique de gestion des données définie à l'échelle de l'établissement. La contribution de toutes les parties prenantes - à commencer par les chercheurs - rassemblées dans un groupe de réflexion *ad hoc* est indispensable. Ce groupe de travail identifiera probablement quelques situations récurrentes pour lesquelles se dessineront des règles simples. Par exemple, il statuera sur la durée de rétention des données fraîchement produites avant la publication du ou des articles correspondants. À l'évidence, il sera saisi de questions plus complexes qu'il instruira en se faisant aider, si besoin, par des experts du domaine. En réponse aux saisines, le Comité conseillera, soit la mise en ligne libre de droits, soit au contraire de ne pas rendre public tel jeu de données, soit de ne rendre les données publiques que sous condition(s). Dans ce dernier cas, il paraît raisonnable d'identifier le demandeur d'accès et de faire préciser l'objectif de la réutilisation, ainsi que d'évaluer les moyens mis en œuvre dans ce but. Les dispositions prises par le demandeur pour valoriser le jeu de données devront également être précisées. La requête, établie selon un dossier type et dûment justifiée, sera analysée par le groupe de travail et aboutira éventuellement à la transmission du jeu de données selon une licence définissant le cadre de son exploitation.
L'enjeu est donc celui d'une véritable expertise collective à construire à l'échelle de l'institution.

5 ■ Recommandations

La réflexion du Comité d'éthique a été éclairée par les travaux réalisés depuis 2013 au sein des deux établissements (chantier sur la gestion et le partage des données pour l'Inra et patrimoine numérique pour le Cirad), ainsi que par l'avis émis par le Comité d'éthique du CNRS. Les recommandations qui en résultent ont pour objet d'identifier les différentes initiatives et actions qui permettraient de répondre au mieux aux questions éthiques soulevées par le partage des données. L'avis du Comité d'éthique suppose que toutes les conditions techniques nécessaires à la gestion et au partage des données, notamment en termes de sécurité, soient acquises.

Rappel de principes

La recherche a pour objectif premier de produire des données dont le traitement et l'interprétation peuvent conduire à de nouvelles connaissances. Cependant, les données issues de la recherche ne peuvent être partagées sans discernement, pour des raisons détaillées dans l'avis du Comité d'éthique. Compte tenu de l'extrême diversité des activités, mais aussi des conditions d'acquisition des données de la recherche à l'Inra et au Cirad, du vaste éventail de possibilités qui en résulte, allant de la mise en ligne sans restriction à la conservation hors ligne, en passant par le partage des données selon des conditions précisément définies, il apparaît indispensable d'envisager un processus de décision applicable au cas par cas. De plus, le partage de jeux de données devra garantir une traçabilité de l'utilisation des données, aussi bien pour les chercheurs des organismes concernés que pour les potentiels utilisateurs extérieurs.
La responsabilité des choix à opérer ne peut relever des seules compétences techniques des responsables des banques de données et, pour des raisons de cohérence, ne peut être laissée à la seule charge des chercheurs.

Recommandations

•1 Le Comité d'éthique ne peut que conforter l'Inra et le Cirad dans leur volonté de partager les données issues de leurs recherches. Le Comité d'éthique recommande de définir une politique d'établissement concernant la gestion et le partage des données qui affiche clairement les engagements des organismes et précise les rôles et responsabilités des différents acteurs. La politique d'établissement devra prévoir un important volet d'information et de formation continue des chercheurs à la mise en ligne des données et à leur réutilisation responsable.

•2 Il paraît essentiel que la définition de cette politique d'établissement soit étroitement concertée avec celle des autres organismes français de recherche publique (CNRS, Inserm, IRD, Ifremer…) et des universités, afin d'aboutir à une réelle mutualisation d'expertise.

•3 Il convient de garantir l'harmonisation de la sémantique, la standardisation des formats, des métadonnées, du contrôle qualité, etc. Ce travail devra être mené notamment au sein d'ELIXIR, de *Research Data Alliance* ou de GODAN.

•4 La politique d'établissement doit aussi évaluer précisément et dégager les moyens humains et techniques nécessaires pour en assurer le succès. En particulier, s'imposent la création et la reconnaissance de nouveaux métiers spécialisés dans le contrôle qualité, l'édition, la gestion et la mise en ligne de données (conservateurs de données), ainsi que dans l'exploitation de ces données (fouille de textes et de données, FTD).

•5 La qualité des données constitue un aspect essentiel de leur fiabilité. Les chercheurs devront y consacrer une attention particulière, mais les organismes devront renforcer et généraliser les procédures de contrôle de cette qualité, ce qui pourra nécessiter des actions spécifiques de formation.

•6 Les activités de production et de partage des données de qualité devront faire l'objet d'une reconnaissance des institutions, et être prises en considération dans l'évaluation des chercheurs, des équipes, unités et organismes.

•7 Les chercheurs travaillant avec leurs partenaires de pays n'ayant pas une politique de gestion des données suffisamment avancée s'attacheront à assurer les formations pertinentes et à veiller à appliquer les standards de protection des données les plus contraignants.

•8 Face à la complexité du sujet, à la diversité des choix possibles et à l'exigence de cohérence à long terme, le Comité d'éthique estime nécessaire de créer une instance spécialisée, traitant des questions soulevées par la gestion et le partage des données, instance qui aura un rôle d'arbitrage (au cas par cas) et qui établira progressivement une jurisprudence d'où les principales règles concernant le partage des données pourront être tirées (commission des ressources numériques, Corenum).

Son rôle implique la présence de chercheurs, mais sa composition devra être beaucoup plus large, incluant les « conservateurs de données » et des personnes à même d'apprécier les interactions avec les aspects non scientifiques, notamment éthiques, administratifs, juridiques, financiers et économiques.
Cette instance contribuera à la définition de la politique des établissements, à l'élaboration de règles concernant le partage des données et à leur nécessaire actualisation en fonction des évolutions du domaine. Cette instance évaluera également le caractère sensible des données (données personnelles, données relatives à la sécurité collective, données concernant la santé ou l'environnement…).

Elle jouera un rôle de référent en matière de gestion et de partage des données pour les chercheurs et les directions de département.

•9 Cette instance évaluera, au cas par cas et en accord avec les groupes de chercheurs ou les chercheurs concernés, si elles peuvent être partagées et, le cas échéant, sous quelles conditions.

•10 En cas de demande de partage de données pour un usage commercial, cette instance veillera, en lien avec l'administration de l'institut concerné, à ce que leur utilisation fasse l'objet de conditions d'usage précises, définies dans un contrat garantissant, d'une part, les droits du fournisseur de données (établissement) et, d'autre part, la valorisation du jeu de données (chercheur et collaborateurs).

Annexes

Annexe 1

Texte de la saisine du Comité : enjeux éthiques de la gestion et du partage des données de la recherche

La multiplication accélérée des données produites et le développement d'outils informatiques permettant de les analyser ouvrent d'importantes perspectives d'application dans nombre de domaines, ce qui n'est cependant pas sans poser des questions d'ordre à la fois politique – voire géopolitique – , économique, juridique et financier aux sociétés. Ces difficultés sont amplifiées dans les pays du sud, peu équipés en lois et règles de protection des données personnelles, mais aussi en moyens d'archivage et de traitement intensif de données numériques, de sorte que l'on évoque parfois l'idée de « double fracture numérique ».

Dans les sciences du vivant, la génomique représente très certainement l'illustration la plus remarquable du « *Big Data* ». Alors qu'il aura fallu plus de 10 ans et quelques milliards de dollars pour réaliser le premier séquençage complet du génome humain (3Gb), il est maintenant possible, à peine une dizaine d'années plus tard, de séquencer l'équivalent de 100 génomes humains sur un seul appareil en quelques jours et pour quelques milliers de dollars. L'automatisation des collectes de données et la numérisation conduisent à la remise en cause d'organisations, pourtant bien en place, quant à leur capacité à archiver, gérer, exploiter et diffuser de telles quantités de données.

Alors que les enjeux technologiques du « *Big Data* » visent à fournir les supports et outils permettant de collecter, stocker et diffuser des masses de données gigantesques, ils encouragent dans le même temps la mise en réseau des citoyens, des chercheurs, des laboratoires et des institutions scientifiques, dans l'optique de développer un vaste réseau d'infrastructures et de collaborations.
Les enjeux scientifiques emportent des interrogations sur la place et les méthodes associées aux démarches guidées par les données, génératrices d'hypothèses, s'appuyant sur des outils permettant d'exploiter, d'intégrer, d'analyser et d'extraire de ces masses de données, souvent hétérogènes et non structurées, des connaissances nouvelles, impossibles à produire autrement, ce qui pourrait conférer aux données une valeur inestimable.

L'« *open science* » recouvre le libre accès, pour tous, aux publications scientifiques et aux données de recherche, mais aussi aux sciences participatives, collaboratives, citoyennes[40], dès lors que cette recherche est financée sur fonds publics, du moins en partie[41]. L'ambition est la validation, la reproduction, voire l'amélioration des résultats de la recherche, l'accélération des progrès scientifiques par une meilleure gestion/documentation des données en vue de leur réutilisation par d'autres. Selon cette approche, l'accroissement de la visibilité et de la citabilité de ces données, du fait même de leur partage, induira des économies par la non-reproduction de travaux de recherche, mais également l'intensification des collaborations dans un cadre mutualisant production de données, méthodes et outils de la recherche. La donnée pourrait acquérir aussi de la valeur en tant qu'un produit de la recherche, référençable, citable et évaluable.

Dans certains domaines des sciences du vivant (génomique, protéomique...), de la physique des hautes énergies et de l'astronomie, des pratiques existent depuis plusieurs années pour assurer la mise en commun, l'archivage et la distribution des données. Il est cependant remarquable que l'essor et le caractère « exemplaire » des entrepôts de données dans le domaine des sciences du vivant ne reposent pas uniquement sur un effort collectif des communautés scientifiques concernées, mais sur la volonté de chaque scientifique de partager ses données. Les créateurs des premiers entrepôts internationaux de données ont pensé que l'idéal d'un partage communautaire suffirait à assurer le dépôt de données dans ces entrepôts. Par crainte de pillage, ce ne fut pas le cas, et pour résoudre le problème, il a fallu imaginer, avec l'appui des journaux scientifiques, des mécanismes permettant de rendre obligatoire le dépôt de données pour pouvoir publier. Ces mécanismes ont posé les bases du fonctionnement actuel qui régit le partage des données dans le domaine des sciences du vivant, ainsi que le modèle économique des éditeurs scientifiques. Il est cependant notable que la massification des données remet régulièrement en cause le modèle de fonctionnement actuel de la recherche.

40 Ces notions ne sont pas équivalentes.

41 Nombre de situations sont à analyser, notamment les données obtenues dans le cadre de partenariat public-privé et/ou dans des contextes légaux qui peuvent différer considérablement d'un pays à l'autre.

La commission européenne étend aujourd'hui les objectifs en matière de libre accès des travaux de recherche qu'elle financera dans le cadre d'« Horizon 2020 ». Les actions engagées et soutenues sont celles favorisant : (1) le libre accès aux données de la recherche (résultats d'expériences, observations, informations produites par ordinateur, etc.) et la mise en place, à titre expérimental, d'un cadre qui tienne compte des questions liées au respect de la vie privée, aux intérêts commerciaux et aux gros volumes de données ; (2) le développement et le soutien des infrastructures interopérables aux échelles européenne et mondiale, pour l'hébergement et le partage des informations scientifiques (publications et données) et (3) l'aide au chercheur qui souhaite mettre ses données en libre accès. Ces propositions de libéralisation constituent une inflexion forte par rapport aux pratiques conventionnelles de la science : articles publiés dans des revues payantes, données partagées à l'initiative d'un individu (et non d'une communauté), données protégées ou secrètes, exclusion des non-pairs...

Confronté aux changements induits par ces évolutions, les chercheurs et leurs institutions doivent faire face de manière croissante à des questionnements, parfois conflictuels, sur leurs responsabilités et obligations (morales, légales) vis-à-vis des pairs, mais aussi du citoyen, des partenaires des projets et des institutions qui financent la recherche.

Si le partage des données de la recherche devient une règle générale, notamment sous l'impulsion des agences de financement, il devient aussi nécessaire d'envisager les questions éthiques liées au partage en complément des questions de propriété intellectuelle[42].

Devons-nous, dans ces conditions, partager TOUTES les données produites, et à quel niveau d'agrégation[43] ? Comment définir la valeur économique et concurrentielle d'une donnée ? Comment préserver l'intérêt des producteurs de données ? Comment établir le caractère « sensible » d'une donnée ? Quelles responsabilités, quelles obligations vis-à-vis des collaborateurs, de la communauté scientifique, de l'institution d'origine ? Comment partager ces données en préservant les intérêts de la science, de la société et du chercheur lui-même ? Existe-t-il des spécificités disciplinaires ou partenariales propres à l'Inra et/ou au Cirad ?

Le partage des données, dans un contexte de « science ouverte » (*open science*), peut-il impliquer un changement de paradigme : passer d'une science pilotée par les hypothèses (*hypothesis-driven science*) à une science pilotée par les données (*data-driven science*). Dès lors, n'y aurait-il pas des risques de concevoir des projets de recherche en fonction des données disponibles et non en fonction de la pertinence de la problématique scientifique ? Comment faire un usage intègre des données produites par d'autres ?

N'y a-t-il pas, au contraire, nécessité d'articuler (et comment) ces deux démarches ? Pour éviter de céder à une recherche dirigée par les données (*data-driven*), **quelle part donner/préserver à la réflexion conceptuelle dans le métier de chercheur ?** Comment et jusqu'où, dans ces conditions, documenter/tracer/donner de la valeur, de la confiance à ces données et qualifier les données des autres ? Comment prendre en compte les données produites et partagées dans l'évaluation des individus ou des collectifs[44] ?

Dans un contexte où la recherche publique est considérée comme une source potentielle de croissance[45] économique, comment les établissements de recherche comme l'Inra ou le Cirad peuvent-ils contribuer au mieux au développement de l'innovation ? Faut-il breveter ou diffuser nos résultats en *open access*[46] ?

Plus récemment, apparaissent des publications scientifiques qui résultent de la participation plus ou moins importante de contributeurs extérieurs au monde académique[47]. Si l'on veut permettre aux citoyens et autres chercheurs de participer à la collecte/l'exploitation des données scientifiques d'un domaine en multipliant les recherches participatives (*crowdsourcing*), il est essentiel de rendre les données scientifiques accumulées compréhensibles et partageables par le plus grand nombre. Quelle responsabilité, quelle reconnaissance et quel rôle pour le chercheur ? Quelle responsabilité, quelle reconnaissance et quel rôle pour le

42 http://www.ehjournal.net/content/10/1/107

43 Le choix du niveau d'agrégation est un point essentiel.

44 http://am.ascb.org/dora/

45 http://www.oecd.org/sti/sci-tech/commercialising-public-research.htm

46 http://www.michaeleisen.org/blog/?p=1301

47 http://dx.plos.org/10.1371/journal.pmed.1001328

bénévole et les collaborateurs ? Ce mode de production de connaissances présente-t-il des dangers et des avantages spécifiques quant à la qualité des données collectées, des résultats publiés, à la propriété des données et des résultats ? À qui appartiennent les données dont la production a été financée tout ou partie sur fonds publics ? Quel positionnement pour les institutions ?

Enfin, dans ce contexte nouveau où les données impulsent de nouvelles pratiques, mais aussi de nouvelles responsabilités, se pose la question du partenariat avec certains pays émergents ou en voie de développement, qui sont fournisseurs et copropriétaires de données, mais souvent ne disposent pas des moyens de traitement et de croisement de données massives.

Les directions des deux établissements s'interrogent aussi sur la démarche éthique et déontologique à mettre en place pour assumer cette nouvelle forme indicible de « pouvoir ».

Annexe 2

Annexe 2 a : Types de données et standards correspondants (Inra)

Extrait de : « Partage des données relatives aux ressources génétiques et génomiques
État des lieux, analyse stratégique et besoins d'accompagnement »

Obtention	Nature	Format standard pour l'échange
brute	séquences lues ADN-ARN	fastq, sff
brute	génotypes SNP	Illumina (matrice marker*organisme) + métadonnées en en-tête, VCF
brute	génotypes SSR	Csv – excel (GeneMapper)
brute	données d'expression (arrays, qPCR)	MIAME
brute	données métabolome	EC number, SBML
brute	profils protéiques (quantitatifs)	?
élaborée	séquences protéines	fasta, asn,embl
élaborée	séquences alignées/assemblées	fasta, bam, sam
élaborée	données d'expression RNAseq	bam, sam, sff,bed, wig
élaborée	polymorphismes SNP	VCF, fasta, flatfile
élaborée	polymorphismes SSR	gff3
élaborée	variants structuraux	VCF (V4.1+)
élaborée	patrons de méthylation	bed ?
élaborée	annotations des gènes	gff3, asn, embl
élaborée	orthologues, paralogues, familles de gènes	Tables xml
élaborée	cartes (génétiques, QTLs, physiques)	acp, text formatted
élaborée	données passeport populations/souches	Voir bases FAO
élaborée	données passeport croisements temporaires	?
élaborée	données passeport banques génomiques	?

Annexe 2 b : Types de données et standards correspondants (Cirad)

Famille(s) de la ressource
OMIQ : Données Omiques
RGB : Ressources Génétiques/Collections biologiques
EXP : Essais/Expérimentations/Analyses laboratoire (mesures instruments...)
OBS : Observations de terrain (relevés, photos, imagerie satellite, film...)
ENQ : Enquêtes (socio-éco, épidémiologie, zootechnie, pratiques culturales...)
TMPS : Séries temporelles (cohortes + séries chrono + mesures répétées)
TAXO : Taxonomie/Ontologie/Référentiel
LOGI : Informatique scientifique originale (logiciels, modèles, scripts,...)
AUTR : Autres (non classable)

Type de données / Jeu de données, Base de données, Carte, Photo, Image, Audio, Vidéo, Logiciel, Service Portail, Web, Objet physique, autres…

Le comite d'éthique : missions et composition

Par décision du 31 octobre 2007, le Cirad et l'Inra ont créé un **Comité consultatif commun d'éthique pour la recherche agronomique**. Ce Comité est placé auprès des Présidents des deux Instituts et a une mission de réflexion, de conseil, de sensibilisation et, au besoin, d'alerte.

Il examine les questions éthiques que peuvent soulever l'activité et le processus de recherche, en France et hors de France, dans les domaines de l'agriculture, de l'alimentation, de l'environnement et du développement durable, et notamment celles qui intéressent les relations entre sciences et société. Le Comité tient compte, en tant que de besoin, des missions et des activités spécifiques des deux Instituts, notamment en matière de recherche pour le développement des pays du Sud. Il peut également conseiller les directions générales des deux établissements pour la mise en place de procédures internes nécessaires à l'application de recommandations formulées par d'autres comités extérieurs institués au plan national, européen ou international, et des réglementations en vigueur relatives à l'exercice de certaines de leurs activités de recherche, en France et hors de France.

Ce Comité commun répond à la logique d'un rapprochement de l'Inra et du Cirad, visant à élaborer une vision partagée des enjeux scientifiques, mondiaux et nationaux, de l'agriculture et de la gestion des ressources vivantes.

Pour l'Inra, ce Comité fait suite au Comepra (Comité d'éthique et de précaution pour les applications de la recherche agronomique), commun à l'Inra et à l'Ifremer (1998-2007). Pour le Cirad, ce nouveau Comité d'éthique fait suite à celui qui avait été mis en place en 2001 et qui avait achevé son mandat en 2005.

Le Comité est présidé par Monsieur Louis Schweitzer.

Il est composé* actuellement de 13 membres :

- Madame **Fifi Benaboud**, Centre Nord-Sud du Conseil de l'Europe,
- Madame **Soraya Duboc**, ingénieur agroalimentaire,
- Madame **Françoise Gaill**, conseillère scientifique à l'Institut écologie et environnement du CNRS,
- Madame **Catherine Larrère**, professeur de philosophie à l'Université Paris 1,
- Madame **Sandra Laugier**, professeur de philosophie à l'Université Paris 1,
- Madame **Jeanne-Marie Parly**, professeur en sciences économiques,
- Monsieur **Jean-Louis Bresson**, médecin nutritionniste et professeur à l'Université Paris 5,
- Monsieur **Marcel Bursztyn**, professeur au Centre pour le développement durable à l'Université de Brasilia (Brésil),
- Monsieur **Paul Clavier**, maître de conférences en philosophie à l'École normale supérieure, Paris,
- Monsieur **Patrick du Jardin**, professeur à Gembloux Agro-Bio Tech, Université de Liège (Belgique),
- Monsieur **Hervé Théry**, géographe, professeur associé à l'Université de Sao Paulo (Brésil),
- Monsieur **Gérard Toulouse**, directeur de recherche au laboratoire de Physique théorique de l'École normale supérieure, Paris,
- Monsieur **Dominique Vermersch**, agronome, professeur d'économie publique et d'éthique, recteur de l'Université catholique de l'Ouest.

* Composition au 16 décembre 2015. Le Conseil d'administration, lors des réunions du 25 mars 2014 et du 25 juin 2014, a nommé trois nouveaux membres (en remplacement de Claude Chéreau, décédé en avril 2014, de Lazare Poamé et Pierre-Henri Tavoillot qui ont souhaité interrompre leur mandat pour raisons personnelles) : Sandra Laugier, Paul Clavier, Hervé Théry.

Les principes et valeurs du Comité d'éthique pour la recherche agronomique

1• Le Comité commun d'éthique considère la reconnaissance de la dignité humaine comme valeur fondamentale. Il s'attachera dans ses recommandations à en donner une application concrète, mettant en œuvre les droits rappelés dans la Déclaration universelle des droits de l'Homme de 1948.

2• Plus généralement, le Comité considère que les valeurs du corpus de déclarations et conventions édifié depuis plusieurs décennies par l'Organisation des Nations unies et les organisations spécialisées, notamment l'UNESCO, font partie de son cadre de référence, parmi lesquelles la protection et la promotion des expressions culturelles, et la biodiversité. La mise en œuvre de ce corpus passe par des accords internationaux normatifs.

3• Il ne faut pas dégrader l'environnement de vie pour les générations futures et ne pas hypothéquer l'avenir de façon irréparable, notamment en épuisant les ressources naturelles ou en mettant en cause les équilibres naturels. Un tel principe de développement durable, impose au Comité de travailler sur le long et le très long terme, et pas seulement sur le court terme. En revanche, le principe d'une réversibilité totale paraît utopique et impraticable.

4• Le monde constitue un système. Toute action sur l'un de ses éléments a des impacts sur d'autres éléments : l'analyse doit alors explorer les effets seconds et induits d'une action et les dynamiques et stratégies qu'elle peut susciter ou favoriser. Les problèmes doivent donc être traités de façon privilégiée à l'échelle mondiale, tout en assurant néanmoins la compatibilité entre le global et le local et en prenant en compte les réalités de terrain.

5• Le Comité considère que la robustesse et l'adaptabilité d'un système sont des éléments positifs. Ainsi, même dans une société ouverte, une certaine autosuffisance dans le domaine alimentaire est souhaitable au niveau national et régional.

6• Le progrès implique une société ouverte aux innovations techniques et sociales, en sachant qu'il faut analyser et prévoir l'impact de ces innovations sur les modes de vie, leur contribution au développement humain, et s'assurer d'un partage équitable des bénéfices qu'elles peuvent apporter.

Coordination : Eric Connehaye • Inra - UCPC | Pierre-Henri Duée • Inra
Conception et réalisation : Studio de création UCPC
Illustration de couverture : ©Inra - Gianpaolo Pagni
Février 2016

Institut National de la Recherche Agronomique (Inra)
147, rue de l'Université 75338 Paris Cedex 07

http://www.inra.fr/l_institut/organisation/l_ethique

Centre de Coopération Internationale en Recherche Agronomique pour le Développement (Cirad)
42, rue Scheffer 75116 Paris

http://www.cirad.fr/qui-sommes-nous/le-cirad-en-bref/notre-organisation/comite-consultatif-commun-d-ethique