

Les données scientifiques face aux enjeux de la recherche en Sciences, Technologie et Médecine : enquête exploratoire à l'Université de Strasbourg

Violaine Rebouillat

▶ **To cite this version:**

Violaine Rebouillat. Les données scientifiques face aux enjeux de la recherche en Sciences, Technologie et Médecine : enquête exploratoire à l'Université de Strasbourg. Études de communication - Langages, information, médiations, 2019, 52, pp.99-116. 10.4000/edc.8681 . hal-02321077

HAL Id: hal-02321077
https://cnam.hal.science/hal-02321077

Submitted on 20 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Les données scientifiques face aux enjeux de la recherche en Sciences, Technologie et Médecine : enquête exploratoire à l'Université de Strasbourg

Violaine REBOUILLAT
Conservatoire National des Arts et Métiers, Laboratoire DICEN-IdF (EA 7339)
violaine.rebouillat.auditeur@lecnam.net

Le 4 juillet 2018, la Ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation, Frédérique Vidal, dévoilait un plan national pour la science ouverte (MESRI, 2018). Doté d'un budget annuel de 3,4 millions d'euros, ce plan entend répondre aux objectifs de l'*Amsterdam Call for Action on Open Science*, à savoir : le libre accès généralisé à l'ensemble des publications scientifiques et la diffusion des données de la recherche financée sur fonds publics. Le Ministère instaure donc une politique en matière de données de recherche, avec pour ambition de systématiser la structuration, la préservation et l'ouverture des données. A cette fin, il s'engage à mettre en place :

- des dispositifs techniques, de type entrepôts de données ;
- des dispositifs financiers, en rendant éligibles les dépenses de traitement des données dans les appels à projet ;
- des dispositifs pédagogiques, telles des formations à la science ouverte dans les écoles doctorales ;
- et des dispositifs administratifs, via la création d'un réseau de correspondants, chargés de répondre aux questions des chercheurs et placés sous la coordination d'un administrateur des données.

Ces différentes propositions reflètent une logique descendante dans l'instauration des pratiques de gestion des données – logique dont on est en droit de questionner l'efficacité. Un processus de co-construction avec les communautés scientifiques ne serait-il pas préférable (Chartron, 2018) ? En tant que principaux acteurs concernés, les chercheurs paraissent en effet être les mieux placés pour déterminer le potentiel des données qu'ils génèrent ou utilisent. Une co-construction des politiques et des dispositifs d'ouverture des données garantirait par ailleurs une meilleure appropriation par les chercheurs.

Dans cet article, nous nous intéresserons à la place qu'occupent les données dans les pratiques de recherche. Nous étudierons en particulier le domaine des Sciences, Technologie et Médecine (STM), en nous appuyant sur six entretiens menés auprès de chercheurs de l'Université de Strasbourg. Ces entretiens font partie d'une enquête plus large, portant sur un échantillon pluridisciplinaire d'une trentaine de projets de recherche. L'enquête s'insère dans le cadre d'une thèse en sciences de l'information et de la communication, dont l'objet d'étude porte sur les données scientifiques et la représentation qu'en ont les différents acteurs impliqués dans leur gestion.

Nous étudierons ainsi dans quelle mesure les stratégies de recherche influencent la gestion et l'ouverture des données en STM.

Nous présenterons tout d'abord les problématiques liées aux données scientifiques en regard des dynamiques de recherche actuelles en STM. Nous exposerons ensuite notre méthodologie d'enquête auprès des chercheurs de l'Université de Strasbourg. Nous montrerons alors en quoi l'objectif d'un projet de recherche influence la manière dont sont gérées les données.

1. Stratégies de la recherche en STM et données scientifiques : dynamiques et projections

Malgré les possibilités qu'ouvrent les technologies numériques de l'internet et du web, l'article de revue reste, dans le domaine des STM, le modèle prédominant en matière de communication scientifique. La science, telle que représentée par ce modèle, est une science qui se lit et s'écrit. Une science dans laquelle le chercheur se considère comme un « producteur de faits » et cherche à convaincre ses pairs que ses énoncés sont des faits. C'est ce qu'observent les sociologues Bruno Latour et Steeve Wooglar au milieu des années 1970, pendant les deux années qu'ils passent en immersion dans un laboratoire de neuroendocrinologie de San Diego en Californie (Latour et Woolgar, 1979). Or, si les chercheurs estiment être des producteurs de faits, pourquoi ne donnent-ils pas davantage de visibilité aux données de recherche, qui sont la preuve de leur raisonnement scientifique ? Probablement, en partie, parce qu'ils sont devenus tributaires d'un système, où la publication dans des revues prestigieuses conditionne leur réussite professionnelle et concentre leur attention. Latour (2001) associe la publication d'articles à un investissement réalisé par le chercheur, dans la perspective de gagner en « crédibilité ». La crédibilité scientifique recouvre les bourses, les postes, les titres honorifiques, mais aussi la confiance accordée par les pairs et la bonne réputation auprès des organismes de financement. Elle est conçue pour être réinvestie, comme le montre le cycle de crédibilité imaginé par Latour (figure 1) : la publication d'articles permet d'obtenir de nouvelles subventions, qui seront investies dans l'achat de matériel ; celui-ci permettra de générer de nouvelles données, grâce auxquelles pourront être formulées de nouvelles théories, qui elles-mêmes seront publiées dans un article, et ainsi de suite. Ce cycle de crédibilité rend compte des mécanismes permettant à un scientifique de continuer à faire des recherches. Il explique à la fois la logique d'ensemble du développement scientifique et les comportements des chercheurs.

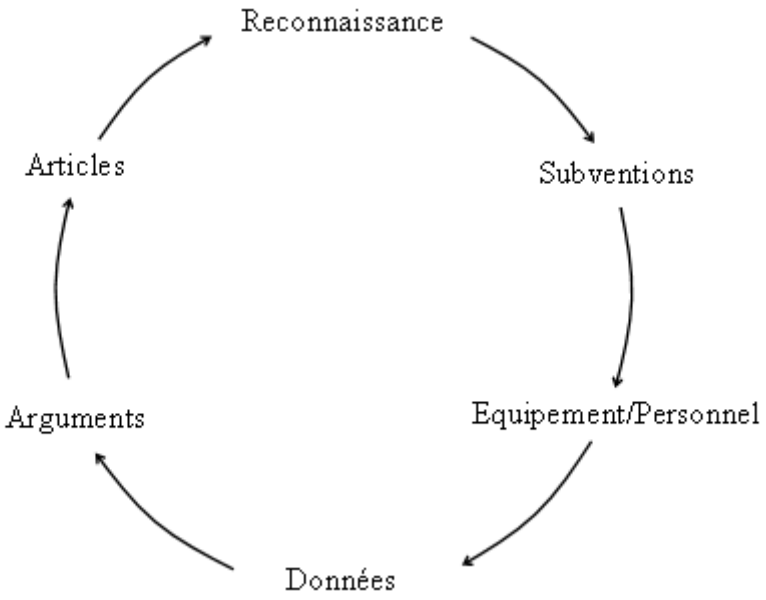


Figure 1 : *Cycle de crédibilité selon Latour (2001, 34)*

Fecher et al. (2015) qualifient le système d'échange de l'information scientifique d' « économie de la réputation » (*reputation economy*). La production scientifique n'est

partagée que si elle permet au chercheur qui la transmet d'accroître sa réputation auprès des pairs. Dans le domaine des STM, la réputation d'un chercheur étant principalement liée au nombre d'articles qu'il publie dans des revues à haut facteur d'impact, celui-ci ne trouvera que peu d'intérêt à diffuser d'autres produits de ses recherches – les données scientifiques en particulier.

A l'ère de l'*openess*, il existe pourtant bel et bien un intérêt pour les données de la recherche et leur ouverture. Certains éditeurs, comme *Nature* et *Science*, demandent de rendre accessibles les données sur lesquelles reposent les conclusions des articles publiés. La mise à disposition des données permettrait en effet la vérification et la reproductibilité des résultats de recherche. Cette règle a été établie pour répondre aux dénonciations régulières de cas de fraude scientifique. Elle reflète donc davantage une préoccupation de préserver la crédibilité de la revue que celle d'offrir une plus grande visibilité aux données de la recherche. En Europe, l'ouverture des données fait partie des lignes politiques directrices de la Commission européenne, comme le rappelle la recommandation du 25 avril 2018 relative à l'accès aux informations scientifiques et à leur conservation (Commission européenne, 2018). La Commission place en effet de nombreuses attentes dans les données de la recherche. Leur ouverture est censée permettre : de reproduire les résultats de recherche, afin de limiter la fraude ; d'accélérer le progrès scientifique, en contribuant à une science plus collaborative ; de stimuler l'innovation, en permettant l'accès et la réutilisation des données par la société civile. Toutefois, comment répondre à ces attentes en termes d'infrastructures ? La question première de l'ouverture des données de la recherche est contenue dans l'objet lui-même : qu'est-ce qu'une donnée de recherche ? Les diverses tentatives de définition ont montré la difficulté à circonscrire cet objet (OCDE, 2007, 18 ; Research Councils UK, 2016 ; Australian National Data Service, 2017). Car la donnée est dépendante du contexte dans lequel elle est utilisée. Selon Borgman, une donnée ne naît pas « donnée » mais le devient. Une entité devient donnée scientifique, lorsque quelqu'un l'utilise comme preuve d'un phénomène dans un but de recherche, c'est-à-dire qu'il la collecte, l'analyse et l'interprète (Borgman, 2015). Schöpfel et al. (2017) envisagent une définition quadridimensionnelle, qui tiendrait compte à la fois de la nature de la donnée, de la communauté disciplinaire dans laquelle elle est utilisée, de son niveau d'enregistrement et de sa finalité. A cette première complexité s'ajoute la question de l'ouverture. L'ouverture ou « libre accès » suppose de rendre les données accessibles à tous (si possible gratuitement) et réutilisables par tous (Open Knowledge International, 2015). Que faire alors des données qui, notamment dans le domaine des STM, présentent des enjeux économiques et sont, à ce titre, considérées comme confidentielles ? Quelle place accorder à l'ouverture des données, quand la priorité est à la publication ?

2. Stratégies de la recherche en STM et données scientifiques : observer les pratiques à l'Université de Strasbourg

2.1. Méthodologie de l'enquête

Pour tenter de répondre à nos questionnements, nous étudierons six projets de recherche, menés dans six laboratoires différents de STM. Notre analyse s'appuie sur les entretiens semi-directifs que nous avons menés avec le coordinateur de chacun des six projets de recherche.

Ces entretiens font partie d'une enquête plus large, menée à l'Université de Strasbourg auprès de chercheurs issus de toutes disciplines. Cette enquête est encore en cours à l'heure actuelle. Nous exposons ici la méthode d'échantillonnage utilisée pour l'enquête, car elle permet de comprendre comment ont été conçus les six entretiens que nous allons analyser dans cet article. Notre échantillon est constitué de 28 projets de recherche, sélectionnés sur la base de

l'excellence scientifique. Nous faisons en effet l'hypothèse que les données d'un laboratoire sont d'autant plus attractives que celui-ci est reconnu par ses pairs. Nous nous sommes donc appuyés sur les rapports d'évaluation 2011-2012 de l'AERES[1] et avons sélectionné les laboratoires de l'Université de Strasbourg ayant obtenu une notation supérieure ou égale à A. Nous avons ainsi obtenu un total de 26 unités de recherche. Dans chacune d'elle, nous avons alors ciblé un, voire deux projets de recherche, ayant cours au moment de l'échantillonnage (soit en novembre 2017) et bénéficiant d'un financement public (type H2020[2] ou ANR[3]). Nous avons ensuite pris contact avec les coordinateurs des différents projets identifiés et leur avons proposé de nous rencontrer pour répondre à nos questions. D'une durée d'environ une heure, les entretiens étaient conduits de manière semi-directive : nous posions au coordinateur scientifique des questions ouvertes, relatives au contexte du projet de recherche, au mode de gestion des données, à leur valeur et leur potentiel partage. Nous avons choisi de mener des entretiens plutôt qu'une enquête en ligne sous forme de questionnaire, afin d'être au plus près du chercheur. Cela nous a notamment permis d'identifier les moments où s'introduisait une confusion entre données et publication – il arrive en effet que les chercheurs identifient les données de recherche aux publications, comme l'avait relevé Francisca Cabrera dans le cadre des entretiens qu'elle a menés en SHS (Cabrera, 2014). En utilisant la forme semi-directive, nous avons par ailleurs pu adopter une approche prospective, qui nous a permis d'isoler des questions de recherche que nous n'avions pas identifiées au préalable. Un des inconvénients des entretiens, que nous avons constatés au cours de l'enquête, est qu'ils conduisent la personne interrogée à tenir un discours plus lissé que si elle avait répondu à un questionnaire en ligne (dans lequel il est plus aisé d'exprimer une opinion crue).

Les six entretiens, que nous avons sélectionnés pour le propos de cet article, ont été réalisés entre novembre 2017 et juin 2018. Leur contenu a été enregistré à l'aide d'un dictaphone puis retranscrit. Les textes ont ensuite été relus attentivement, de manière à en isoler les thèmes récurrents. L'analyse a alors consisté à comparer ces thèmes entre eux et à les confronter aux caractéristiques des différents projets de recherche (discipline, visée du projet, conditions de travail, types de données…).

2.2.Présentation des projets de recherche étudiés

Les six projets de recherche, que nous présenterons ici, ont pour point commun de comporter des enjeux en santé humaine. Ils relèvent de différentes disciplines scientifiques, allant de la chimie à la biologie, mais tous s'attachent à observer, comprendre ou modifier des mécanismes du corps humain.

Le **projet 1** est porté par un laboratoire de neurosciences, auquel sont associés trois autres laboratoires français du même domaine. Il s'agit d'un projet de recherche fondamentale, dont l'objectif est de mieux comprendre le rôle du thalamus, une des structures de notre cerveau, dans le processus de mémoire à long terme. Jusqu'à présent, on n'attribuait en effet au thalamus qu'une simple fonction de transmission de l'information entre le cerveau

---

[1] Rapport d'évaluation des unités de recherche 2011-2012 : https://www.hceres.fr/LISTE-ALPHABETIQUE-DES-ETABLISSEMENTS-ET-ORGANISMES-EVALUES/UNIVERSITE-DE-STRASBOURG-UNISTRA. Il s'agissait de la dernière évaluation rendue publique dans sa totalité, lorsque nous avons constitué notre échantillon en octobre 2017.

[2] Agence Nationale de la Recherche (http://www.agence-nationale-recherche.fr/)

[3] Horizon 2020 est le programme de financement 2014-2020 de la Commission européenne (http://www.horizon2020.gouv.fr/).

et le reste de l'organisme. Les équipes de recherche travaillent sur des modèles animaux (le rat notamment).

Le **projet 2** relève également des neurosciences. Il s'agit d'un projet de recherche fondamentale, mené par une équipe unique et portant sur la sclérose latérale amyotrophique, une maladie neuro-dégénérative qui affecte les neurones impliqués dans les mouvements. C'est une maladie qu'on ne sait pas soigner à l'heure actuelle et qui conduit au décès des patients en trois à cinq ans. L'originalité du projet de recherche est d'étudier l'implication d'une population particulière de neurones dans l'évolution de la maladie. Il s'agit de neurones situés dans le cortex cérébral, peu étudiés jusqu'à présent dans le contexte de la sclérose latérale amyotrophique. Tout comme le projet 1, l'implication de ces neurones est analysée à partir de modèles animaux (des souris).

Le **projet 3** est un projet européen, qui rassemble six partenaires des domaines de la chimie et de la biologie. Le partenaire français est une équipe spécialisée dans la chimie thérapeutique. Il s'agit d'un projet de recherche fondamentale. L'objectif est d'identifier les impacts du graphène sur la santé, en particulier l'éventuel dérèglement des gènes qu'il pourrait entraîner. Le graphène est un nouveau matériau, aujourd'hui considéré comme le matériau du futur pour l'électronique, l'optique ou encore la médecine. L'équipe française est chargée d'élaborer différents composés du graphène, qui seront ensuite testés par les partenaires européens – in vitro, sur des cellules du système immunitaire, et in vivo, sur des modèles animaux.

Le **projet 4** est porté par un laboratoire de chimie thérapeutique, auquel est associée une unité de recherche en neurosciences. Les deux équipes s'intéressent aux douleurs chroniques neuropathiques. Engendrées par une lésion nerveuse, ces douleurs touchent aujourd'hui 15% de la population. Or il n'existe actuellement aucun traitement permettant d'en cibler la cause. Les douleurs chroniques neuropathiques ont donc un coût socio-économique très important en termes d'assurance maladie et de chômage. Le but du projet de recherche est de concevoir une molécule administrable sous forme de comprimé, capable de stopper le mécanisme à l'origine de ces douleurs. Il s'agit d'un projet de recherche appliquée, ayant fait l'objet de plusieurs dépôts de brevets et qui conduira, à terme, à la commercialisation d'un médicament.

Le **projet 5** est coordonné par un chercheur en biologie moléculaire. Il consiste à développer un nouvel outil moléculaire dédié à l'imagerie (on parle de « sonde moléculaire »). L'objectif est de créer une technologie capable d'entrer dans une cellule vivante et d'émettre une fluorescence, afin de pouvoir ensuite suivre par imagerie la synthèse d'ARN dans cette cellule. Il s'agit d'un projet de recherche appliquée, qui aboutira au dépôt de plusieurs brevets. A l'équipe de biologie moléculaire sont associées deux autres équipes : une équipe de chimie, qui contribue au développement de la sonde moléculaire avec l'équipe de biologie ; une équipe de bactériologie, chargée de démontrer la fonctionnalité de la sonde élaborée, en l'appliquant sur des bactéries.

Le **projet 6** s'intéresse aux cellules de Sertoli, situées dans les testicules et responsables de l'organisation des cellules germinales à chaque étape de leur différenciation. Le projet est porté par un laboratoire de génétique, auquel s'associe un laboratoire de biologie, choisi pour son expertise en bioinformatique. L'objectif est de déterminer le rôle de l'acide rétinoïque dans la cellule de Sertoli au niveau de l'expression des gènes. Le projet s'inscrit dans le domaine de la recherche fondamentale.

2.3.Intérêt des chercheurs interrogés pour les données scientifiques

Durant les entretiens, nous avons demandé aux coordinateurs s'ils utilisaient le terme « donnée » dans le cadre de leurs recherches. Tous nous ont répondu par l'affirmative. Le mot « donnée » est en effet plus usuel en STM que dans les sciences humaines par exemple

(Cabrera, 2014, 52 ; Schöpfel, 2018, 10), dans la mesure où son acception renvoie originellement au caractère quantitatif de ce qui a été mesuré, c'est-à-dire à la donnée chiffrée.

Nous avons alors interrogé les six coordinateurs sur le sens qu'ils donnaient à ce terme. Les définitions formulées par chacun sont rassemblées dans le tableau 1. Ces définitions ont pour point commun d'associer les données au résultat de l'expérience. Il y a là un premier aspect lié à la typologie des données : les données produites dans les disciplines en question (chimie, biologie moléculaire, génétique, neurosciences) sont des données d'expérimentation. A la différence des données d'observation, elles ont sciemment été générées par les équipes de recherche, dans un espace maîtrisé qu'est la « paillasse » ou le laboratoire. On peut donc rattacher ce premier aspect à la dimension « nature de la donnée », proposée dans la définition de Schöpfel et al. (2017). Le second aspect se rapporte à la dimension « communauté disciplinaire », dans la mesure où l'expérimentation est une méthodologie de recherche caractéristique de la chimie et de la biologie – ces disciplines visant, par essence, à comprendre ce qui se passe dans la nature, voire à créer à partir d'éléments se trouvant dans la nature.

	Définition de « donnée de recherche »	**Exemples de données générées**
Projet 1	« le produit de l'expérimentation, qu'il soit qualitatif ou quantitatif »	Données de mesure du comportement des rats Données d'histologie
Projet 2	« toute information qui peut conduire à un résultat » « ce qu'on interprète, ce qu'on analyse, ce qu'on manipule »	Données de mesure du comportement des souris Données d'histologie Données de séquençage
Projet 3	« les résultats scientifiques qu'on obtient des différentes expériences faites dans le laboratoire »	Formule chimique du matériau élaboré Données de mesure du matériau
Projet 4	[N'a pas donné de définition]	Données de mesure du comportement des souris Formules chimiques des molécules élaborées
Projet 5	« l'output de l'expérience », « ce que l'expérience va générer, l'information qu'elle va nous donner »	Données de mesure de la fluorescence Données de séquençage
Projet 6	« quelque chose d'informatique », « quelque chose qui a une composante chiffrée »	Données de spectrométrie Données de séquençage

Tableau 1 : *Définitions de « donnée de recherche » proposées par les chercheurs interrogés*

Quant à l'expression « ouverture des données », nous n'avons pas eu de mal à nous faire comprendre en l'utilisant. La question du libre accès est en effet un sujet relativement familier des chercheurs de l'Université de Strasbourg. Le discours politique incitatif de la direction de l'Université y a probablement contribué, notamment depuis la création de l'archive ouverte UnivOAK en 2016. Toutefois, si la publication en libre accès est devenue pour les chercheurs une question d'actualité, qu'ils n'hésitent pas à remettre en cause, l'ouverture des données

reste encore une réalité lointaine. Parmi les chercheurs rencontrés, certains avaient déjà déposé des données en ligne – notamment des codes sources et des données de séquençage (projets 2, 4 et 5) – mais ces dépôts restent à l'état de pratiques ponctuelles, motivées par l'injonction des éditeurs au moment de la publication d'un article. L'ensemble des six chercheurs se sont cependant montrés favorables au principe de mise à disposition des données. Parmi les arguments cités, la lutte contre la fraude scientifique a systématiquement été mentionnée. En revanche, les chercheurs n'ont pas semblé avoir conscience de ce qu'implique la diffusion des données en termes d'infrastructure technique et humaine.

3. De la recherche fondamentale à la recherche appliquée en STM : quelle incidence sur les données ?

3.1. La nécessité d'une recherche subventionnée

Dans les disciplines que nous avons étudiées (chimie, biologie moléculaire, neurosciences), il ne peut y avoir de données sans financement. Ces disciplines relèvent en effet des sciences expérimentales, dans lesquelles la génération de données dépend de matériel souvent coûteux.

Lorsque la recherche passe par l'étude de modèles animaux (souris, rats…), elle nécessite des frais d'animalerie sur une durée relativement longue – l'étude étant soumise au rythme biologique des animaux (croisements, évolution d'une maladie…). De même certains produits de type réactifs, utilisés pour des expériences à la paillasse, peuvent être extrêmement coûteux. En biologie, le séquençage d'ADN ou d'ARN est relativement fréquent et a lui aussi un coût élevé (trois des six projets étudiés y ont recours). A l'Université de Strasbourg, le séquençage est réalisé par une unité de recherche sous forme de prestation de service. Externaliser le séquençage est certes moins coûteux qu'investir dans un séquenceur, mais le montant de la prestation reste tout de même élevé. A titre d'exemple, le budget investi par le projet 2 dans le séquençage d'ADN s'élève à 50 000€. Enfin, à tout cela il faut ajouter le salaire des ingénieurs, techniciens et doctorants chargés de réaliser des expériences souvent longues et minutieuses.

Au vu des budgets requis, on pourrait donc penser que la réutilisation de données est judicieuse. Or celle-ci ne semble pas être une pratique courante dans les domaines rencontrés. A ce jour, nous ne pouvons l'expliquer que par une exigence d'originalité. Comme l'a expliqué l'un des coordinateurs, « on ne peut pas utiliser deux fois les mêmes données ; il faut de nouveau préparer, caractériser un matériel, ou alors faire référence à quelque chose qui a déjà été publié ». Les données utilisées dans un article ne peuvent faire l'objet d'une seconde publication. Cela reflète, d'une part, l'exigence des éditeurs qui, dans le modèle éditorial classique, disposent des droits de diffusion exclusifs sur l'article. D'autre part, cela renvoie à une conception cumulative des sciences, qui se doivent de produire des connaissances toujours inédites.

L'obtention de financements réguliers fait donc partie des principales préoccupations des équipes de recherche. Le rôle d'un chef d'équipe implique d'ailleurs de prendre du temps pour répondre à des appels à projets, afin d'obtenir de nouveaux financements. Car de l'accord de nouvelles subventions dépend l'acquisition de données et, par suite, la possibilité d'aboutir à des résultats de recherche.

3.2.La publication d'abord, les données ensuite

L'analyse des entretiens nous a permis de constater que le mode de fonctionnement des projets de recherche étudiés s'accordait avec le cycle de crédibilité élaboré par Bruno Latour

(figure 1). Dans ces disciplines, en effet, il n'y a pas de financement sans crédit-reconnaissance et pas de reconnaissance sans publication dans une revue scientifique. Or, comme nous l'avons montré précédemment, obtenir un financement est une étape nécessaire à la réalisation d'un projet de recherche. La publication d'articles scientifiques constitue donc un enjeu clef pour obtenir ces financements et faire ainsi en sorte que le travail de recherche se poursuive. Comme nous l'expliquait le coordinateur du projet 1, « si [une équipe de recherche] tombe sur un résultat majeur, [elle va] chercher à le publier dans un journal de forte réputation et, par ce biais-là, [elle va] crédibiliser de futures demandes de financement ». L'article de revue prévaut donc comme structure d'organisation, de validation et de circulation des savoirs scientifiques. C'est lui qui, par sa publication, permet de gagner en crédibilité – en « réputation » (Fecher et al., 2015) – auprès des pairs et des organismes de financements.

Or, selon Fecher et al. (2017), dans ce système de communication de l'information scientifique, donnant la priorité à la publication d'articles, les données de recherche ne possèdent qu'une faible valeur d'échange. D'où un impact sur leur diffusion et leur gestion. Il est en effet rare que les fichiers de données soient conservés sur le long terme, notamment lorsque surviennent des problèmes d'encombrement. Seules les données sous-jacentes à une publication font l'objet d'une conservation plus assidue, avec souvent triple ou quadruple support de sauvegarde. Quant à la diffusion des données, elle ne fait pas l'objet de pratiques spontanées chez les chercheurs que nous avons rencontrés. Nous l'évoquions en fin de deuxième partie. Unanimement, les six coordinateurs voient dans la diffusion des données une valeur d'administration de la preuve. Rendre les données disponibles permet de reproduire un résultat et ainsi de limiter la fraude. Là encore, seules les données sous-jacentes à une publication sont concernées. Cet usage des données comme administration de la preuve va d'ailleurs toujours dans le sens d'un modèle de communication des savoirs fondé sur la place prépondérante de l'article scientifique.

3.3. Vers une valorisation des données dans la recherche appliquée

La recherche appliquée semble se distinguer de la recherche fondamentale dans la manière dont sont gérées les données. Dans les projets de recherche appliquée, en l'occurrence les projets 4 et 5, l'attention portée aux données nous a en effet paru plus grande que dans les projets 1, 2, 3 et 6, relevant de la recherche fondamentale.

Nous commencerons par définir les termes de recherche fondamentale et de recherche appliquée. Pour cela, nous nous appuyons sur les travaux de l'OCDE. Ceux-ci décrivent la recherche fondamentale comme « [consistant] en des travaux expérimentaux ou théoriques entrepris principalement en vue d'acquérir de nouvelles connaissances sur les fondements des phénomènes et des faits observables, sans qu'il y ait une application ou une utilisation particulière en vue ». La recherche appliquée, quant à elle, est définie par l'OCDE comme « [consistant] également en des travaux originaux entrepris en vue d'acquérir des connaissances nouvelles. Cependant, elle est surtout dirigée vers un but ou un objectif pratique déterminé. […] Les connaissances ou les informations tirées de la recherche appliquée sont généralement susceptibles d'être brevetées mais peuvent également être conservées secrètes ». Conscients de la porosité des frontières entre ces deux catégories de recherche, nous ne discuterons toutefois pas ces définitions, car tel n'est pas l'objectif du présent article. Précisons néanmoins que nous n'avons pas nous-mêmes attribué l'étiquette de recherche fondamentale ou appliquée aux différents projets étudiés, mais que ce sont bien les chercheurs interrogés qui ont défini leur projet comme relevant de tel ou tel type de recherche.

Dans la recherche appliquée, l'importance de publier dans des revues à haut facteur d'impact existe bel et bien, mais elle se double d'un souci d'assurer le transfert de la

technologie qui a été créée. Le gain de « crédibilité » passe donc par la publication d'articles (figure 2, cycle 1), mais pas seulement. Selon nous, la crédibilité s'acquiert également par le transfert de technologie, qui est l'enjeu même des recherches appliquées (figure 2, cycle 2). Pour ce type de projets, le cycle de crédibilité serait donc double. La réussite du transfert de technologie apporte en effet des bénéfices financiers aux tutelles, puisque dans le cadre d'une invention les tutelles sont propriétaires du brevet déposé. Grâce à ces bénéfices, les chercheurs s'attirent la confiance des tutelles et indirectement des agences de financement qui, par la suite, leur attribueront d'autant plus volontiers de nouvelles subventions. Dans le cycle 2, la crédibilité provient donc des bénéfices financiers issus du transfert de technologie.

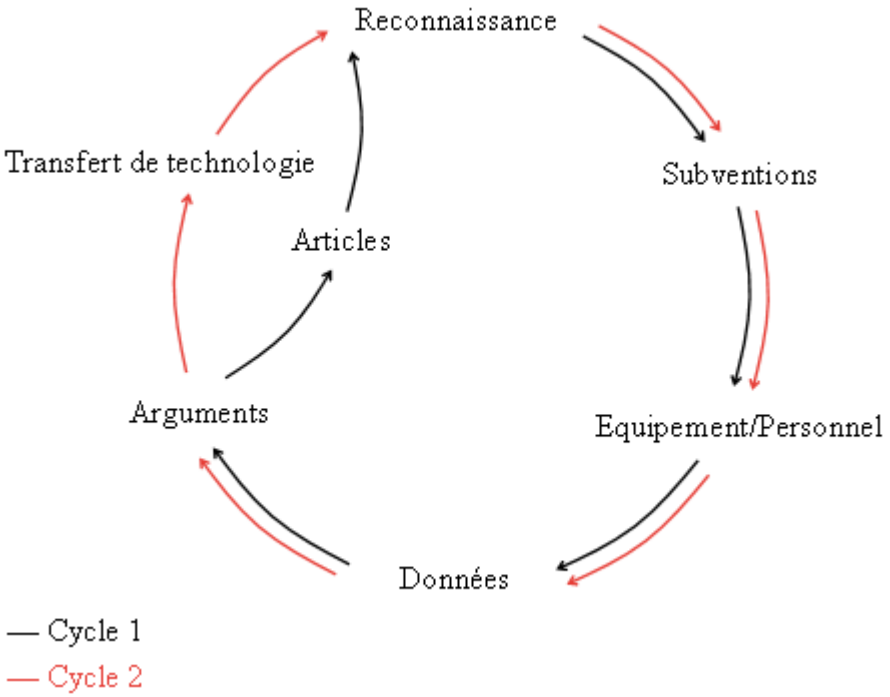


Figure 2 : *Cycles de crédibilité dans la recherche appliquée*

C'est pourquoi, dans un projet de recherche appliquée, obtenir de la crédibilité et, par la suite, de nouveaux financements suppose de mener à bien le transfert de technologie. Or nous avons constaté que celui-ci dépend de la manière dont sont gérées les données de recherche. Le coordinateur du projet 4 nous a expliqué qu'« à partir du moment où on veut valoriser des résultats, où on va trouver un partenaire industriel qui s'intéresse à ces résultats, il va falloir que l'industriel soit convaincu des données que nous avons générées – ce qui suppose de mettre en place un certain nombre de pratiques ». Pour garantir la reprise d'une invention par l'industrie, il est nécessaire que les données aient été (1) validées, (2) conservées et (3) protégées tout au long du projet de recherche. (1) Valider les données suppose de s'être assuré d'un résultat avant d'entamer l'étape suivante de la recherche. Par ailleurs, la tenue rigoureuse des cahiers de laboratoire (que nous avons spécifiquement observée dans les projets 4 et 5) permet de garantir la traçabilité des données. (2) Conserver les données consiste, pour les projets que nous avons étudiés, à stocker l'ensemble des données générées et à les répliquer sur des serveurs sécurisés et géographiquement distants. (3) Enfin, protéger les données vise à préserver leur caractère inédit, c'est-à-dire à garantir l'exclusivité de la découverte et de son exploitation. C'est pourquoi les données, lorsqu'elles ont besoin d'être échangées entre les

partenaires du projet, sont transmises via un réseau de messagerie non commercial voire sous forme de fichiers cryptés. Ces pratiques, qui divergent de la recherche fondamentale telle que nous l'avons observée à l'Université de Strasbourg, témoignent d'une préoccupation plus grande pour les données scientifiques. Car, dans un contexte d'innovation, ces dernières possèdent une valeur économique indirecte. Ayant d'ores et déjà adopté une démarche de gestion rigoureuse des données, les scientifiques menant des recherches appliquées seront donc probablement plus sensibles à la politique de structuration et de préservation des données conduite par la Commission européenne et le Ministère de la Recherche. En revanche, en raison des enjeux économiques sous-jacents, le caractère confidentiel des données produites par de telles recherches rendra très certainement leur ouverture difficile, du moins tant que le processus de protection (souvent long et contraignant) ne sera pas achevé.

4. Conclusion

Comme nous l'avons vu, dans les sciences expérimentales, la collecte et l'analyse de données coûtent cher, rendant indispensable l'obtention de financements. Ce pré-requis conditionne donc fortement les stratégies de recherche. En effet, les chercheurs sont amenés à donner la priorité à ce qui attire la faveur des financeurs, à savoir la publication d'articles dans des revues prestigieuses, les transferts de technologies... Or, nous constatons que cela a une incidence sur la manière dont sont gérées les données scientifiques. Quand, dans la recherche fondamentale, c'est la publication dans des revues prestigieuses qui prime, les chercheurs jugent peu important de sauvegarder ou partager les données. Lorsque, dans la recherche appliquée, il s'agit de mener à bien un transfert de technologie, le souci de préserver les données est plus grand, car de lui dépend la confiance accordée par le partenaire industriel. La gestion et le partage des données en STM dépendent donc fortement de l'objectif du projet de recherche.

Ces conclusions, malgré leur portée limitée, due au petit nombre d'entretiens sur lesquels elles s'appuient, mettent en évidence différentes manières de gérer les données. Indirectement, elles révèlent la diversité des contextes de recherche et leur incidence sur la gestion des données. Une politique publique en matière de structuration et d'ouverture des données scientifiques aura probablement peu d'effets, tant que les spécificités disciplinaires, techniques, socio-économiques… qui constituent le monde de la recherche, n'auront pas été étudiées dans leur diversité. Nous pensons que c'est en s'appuyant sur l'existant et en co-construisant avec les communautés de recherche qu'une telle politique aura le plus de chance d'être adoptée par les scientifiques.

Australian National Data Service (2017). *What is Research Data? ANDS Guide*. Disponible sur https://www.ands.org.au/guides/what-is-research-data (page consultée le 10 septembre 2018).

Borgman C. L. (2015). *Big data, little data, no data: scholarship in the networked world*, Cambridge (Massachusetts), The MIT Press.

Cabrera F. (2014). *Les données de la recherche en Sciences humaines et sociales : enjeux et pratiques. Enquête exploratoire.* Mémoire, CNAM-INTD.

Chartron G. (2018), à paraître. « L'open science au prisme de la Commission européenne ». In *Education et Sociétés*.

Commission européenne (2018). *Recommandation de la Commission du 25.4.2018 relative à l'accès aux informations scientifiques et à leur conservation*. Disponible sur https://ec.europa.eu/transparency/regdoc/rep/3/2018/FR/C-2018-2375-F1-FR-MAIN-PART-1.PDF (page consultée le 10 septembre 2018).

Fecher B., Friesike S., Hebing M., Linek S. (2017). « A reputation economy: how individual reward considerations trump systemic arguments for open access to data ». In *Palgrave Communications*, vol. 3. Disponible sur https://www.nature.com/articles/palcomms201751 (page consulté le 10 septembre 2018). doi:10.1057/palcomms.2017.51

Fecher B., Friesike S., Hebing M., Linek S., Sauermann A. (2015). « A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing ». In *DIW Berlin Discussion Paper*, vol. 1454. Disponible sur https://ssrn.com/abstract=2568693 (page consultée le 10 septembre 2018). doi:10.2139/ssrn.2568693

Latour B. (2001). *Le métier de chercheur : regard d'un anthropologue*. Versailles, Editions Quæ. Disponible sur https://www.cairn.info/le-metier-de-chercheur--9782738009739.htm (page consultée le 10 septembre 2018). doi:10.3917/quae.latou.2001.01

Latour B., Woolgar S. (1979). *La vie de laboratoire. La production des faits scientifiques*, Paris, La Découverte.

Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (2018). *Plan national pour la science ouverte*. Disponible sur http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf (page consultée le 10 septembre 2018).

Open Knowledge International (2015). *Open Definition*, version 2.1. Disponible sur https://opendefinition.org/od/2.1/en/ (page consultée le 10 septembre 2018).

Organisation de Coopération et de Développement Économiques (1975). *La mesure des activités scientifiques en techniques : méthode-type proposée pour les enquêtes sur la recherche et le développement expérimental. « Manuel de Frascati ».* Disponible sur https://hal.archives-ouvertes.fr/hal-01511852 (page consultée le 10 septembre 2018).

Organisation de Coopération et de Développement Economiques (2007). *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*. Paris, Editions OCDE. Disponible sur http://www.oecd.org/fr/science/sci-tech/38500823.pdf (page consultée le 10 septembre 2018).

Research Councils UK (2016). *Concordat on Open Research Data*. Disponible sur https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/ (page consultée le 10 septembre 2018).

Schöpfel J. (2018). *Vers une culture de la donnée en SHS : Une étude à l'Université de Lille*. Villeneuve-d'Ascq, GERiiCo, Université de Lille, rapport de recherche. Disponible sur https://hal.archives-ouvertes.fr/hal-01846849 (page consultée le 10 septembre 2018).

Schöpfel J., Kergosien E., Prost H. (2017). « Pour commencer, pourriez-vous définir « données de la recherche » ? Une tentative de réponse ». Toulouse, INFORSID 2017. Disponible sur https://hal.univ-lille3.fr/hal-01530937 (page consultée le 10 septembre 2018).