Vers une culture de la donnée en SHS

Joachim Schöpfel

Vers une culture de la donnée en SHS.
Une étude à l'Université de Lille

Rapport

Joachim Schöpfel

Université de Lille, laboratoire GERiiCO

« *La science est un bien commun que nous devons partager le plus largement possible.* »

Frédérique Vidal, ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation

Villeneuve d'Ascq, juillet 2018

L'auteur :

Joachim Schöpfel est enseignant-chercheur (MCF-HC) en Sciences de l'Information et de la Communication à l'Université de Lille Sciences Humaines et Sociales et membre du laboratoire GERiiCO. Après avoir dirigé l'UFR IDIST de 2009 à 2012 et l'Atelier National de Reproduction des Thèses de 2012 à 2017, il travaille aujourd'hui comme consultant indépendant dans le domaine de l'information scientifique et technique, est membre d'euroCRIS, de Greynet, de NDLTD et de Knowledge Exchange ; il conduit un projet dans le domaine des données de la recherche et monte actuellement un projet franco-allemand sur l'évolution des thèses de doctorat dans l'environnement de la science ouverte.

ORCID 0000-0002-4000-807X joachim.schopfel@univ-lille.fr

Une étude du projet D4Humanities, financé par la MESHS et par le Conseil Régional Hauts-de-France.

Table des matières

Résumé

La science ouverte figure parmi les priorités de l'Etat français. Dans la continuité des chantiers engagés par le gouvernement français sur la transformation numérique de l'Etat et sa modernisation, le deuxième plan d'action national 2018-2020 *Pour une action publique transparente et collaborative* précise que la France « soutient la mise en œuvre des principes du gouvernement ouvert pour renforcer (…) l'accès aux matériaux et résultats de la recherche ». Le plan national pour la science ouverte, présenté début juillet 2018, a confirmé cette ambition. L'objectif est que les données produites par la recherche publique soient progressivement structurées en conformité avec les principes FAIR, préservées et, quand cela est possible, ouvertes.

Le développement d'un écosystème de la science ouverte s'inscrit dans un contexte mondial, caractérisé par la technologie numérique, les réseaux organisés en ligne et les outils collaboratifs. Aussi, la LERU, la Ligue des Universités de Recherche Européennes, vient de publier une feuille de route (*advice paper*) pour accélérer le développement d'une culture de science ouverte au sein des établissements. La LERU plaide pour un changement culturel et propose un cadre général, à partir des huit axes de la Commission Européenne.

Notre étude *Vers une culture de la donnée en SHS* souhaite contribuer à la mise en œuvre de l'écosystème de la science ouverte sur le terrain d'un campus universitaire. L'étude a été réalisée dans le cadre du projet structurant *D4Humanities*, avec un financement de la MESHS et du Conseil Régional Hauts-de-France, et elle fait suite à des travaux de recherche menés depuis 2013 par le laboratoire GERiiCO. Conduite sous forme d'entretiens avec 51 chercheurs, doctorants, responsables de laboratoires, chefs de projets et ingénieurs en charge de données, l'étude poursuit trois objectifs :

1. (Re)Mettre les enseignants-chercheurs au cœur de la mise en œuvre de l'écosystème de la science ouverte sur le campus, avec leurs besoins, priorités et interrogations.
2. Identifier des opportunités et verrous pour une politique de données.
3. Recommander dix actions à mettre en place pour développer la culture de données sur le campus.

Menée comme un audit sur un terrain particulier et dans le domaine des sciences humaines et sociales, l'étude se réfère à d'autres travaux, dont notamment les enquêtes des universités Rennes 2, Strasbourg, Humboldt de Berlin et Wageningen, les études menées par la Direction IST du CNRS et d'autres recherches.

Notre étude a une portée pragmatique: dégager les éléments indispensables pour une politique cohérente de la production, gestion et réutilisation des données de la recherche sur un campus en sciences humaines et sociales, et contribuer ainsi à l'appropriation du concept de la science ouverte par une « mise en culture de la donnée, qui effectue une mise en sens d'usages disséminés et spécialisés de données ouvertes ».

Une première partie (« Constats préalables ») s'appuie sur deux études (Rennes 2, Lille 3) pour mieux cerner le concept de la donnée de recherche et son caractère de « longue traîne » ; cette partie synthétise les pratiques, motivations et attentes des enseignants-chercheurs dans ce domaine, en SHS. Elle aborde également d'une manière générale la question des services et dispositifs de données.

Une deuxième partie (« Observations ») décrit un paysage contrasté à partir des entretiens menés en 2017 et 2018 sur le campus SHS de l'Université de Lille. Les besoins prioritaires des chercheurs sont la sécurité des données et systèmes, et la communication au sein des projets. L'image qui se dégage est un continuum de pratiques plus ou moins efficaces, formalisées et adéquates, avec une gouvernance

parfois incertaine, au niveau des projets aussi bien qu'au niveau des structures. Ces pratiques sont liées aux communautés disciplinaires mais plus encore, aux méthodes, équipements et thématiques scientifiques. Tous les chercheurs ne parlent pas le même langage quand il s'agit des résultats de leur travail et de la diffusion de ces résultats : ce fait ne facilite pas le développement d'une culture partagée de la donnée. Les incitations pour une telle culture sont multiples : les programmes de recherche européens (H2020) et français (ANR), le règlement général sur la protection des données, les normes éthiques, les politiques éditoriales des revues scientifiques etc. En face, plusieurs verrous, dont avant tout un manque de ressources informatiques, personnelles et financières ; quand il s'agit de bonnes pratiques et de contraintes dues à la législation ou aux programmes de recherche, le verrou majeur n'est pas d'ordre psychologique ou sociologique mais matériel.

La troisième partie (« Vers une culture de la donnée ») liste d'une manière succincte dix recommandations qui, ensemble, définissent un cadre de référence pour la mise en œuvre d'une politique de données sur un campus SHS :

1. Mettre en place un pilotage scientifique
2. Investir d'une manière ciblée
3. Viser les projets, pas les laboratoires
4. Utiliser les plans de gestion comme levier
5. Apporter des réponses aux contraintes de sécurité
6. Apporter des réponses aux besoins de communication
7. Apporter des réponses aux besoins de curation
8. Proposer plusieurs solutions pour la conservation des données
9. Institutionnaliser le lien avec la TGIR Huma-Num
10. Soutenir les bonnes pratiques

L'essentiel de cette proposition tient en trois points :

1. Mettre en place un pilotage scientifique, pour une coordination des actions et services de l'ensemble des acteurs (services centraux et communs, structures de recherche, composantes etc.).
2. Concentrer la politique sur certaines actions ciblées, en tenant compte des priorités des chercheurs et en mettant l'accent sur les projets de recherche (H2020, ANR etc.).
3. Positionner la démarche clairement au sein des infrastructures nationales et européennes en SHS, en particulier par une institutionnalisation des liens avec la TGIR Huma-Num.

En revanche, il faut éviter tout discours d'injonction idéologique sur la question de l'ouverture des données, tout comme il faut éviter l'éparpillement des efforts et ressources. Il n'est pas possible de donner une réponse à toutes les demandes, d'autant qu'une partie des solutions se trouvent à l'extérieur du campus, dans les projets et communautés de recherche, dans les infrastructures et services au plan national, et dans les réseaux internationaux. La politique à mener devrait appliquer des principes de subsidiarité et de complémentarité, ce qui implique une très bonne connaissance du terrain de la recherche et des dispositifs de données.

Deux autres actions devraient accompagner une telle politique de campus :

- La création d'une bibliothèque de référence sur les données et la science ouverte, regroupant des rapports, études, standards, manuels, recommandations, ouvrages de références etc.
- La mise en place d'un projet scientifique multidisciplinaire pour analyser les besoins et pratiques, pour évaluer les politiques, dispositifs et modèles économiques, et pour assurer le suivi de la mise en œuvre d'un écosystème de la science ouverte.

Le plan d'action national 2018-2020 constate qu'il reste « encore beaucoup à faire pour que la science ouverte prenne toute sa place dans les pratiques scientifiques ». Pour réussir, une telle démarche nécessite une connaissance de la réalité du terrain ; elle a besoin du soutien des communautés de chercheurs, d'une coordination de tous les acteurs sur le campus et d'un pilotage institutionnel et scientifique. Cela prendra du temps. Mais c'est un investissement nécessaire pour maintenir l'excellence de la recherche.

Introduction

La science ouverte figure parmi les priorités de l'Etat français. Dans la continuité des chantiers engagés par le gouvernement français sur la transformation numérique de l'Etat et sa modernisation, le deuxième plan d'action national 2018-2020 *Pour une action publique transparente et collaborative*[1] précise que la France « soutient la mise en œuvre des principes du gouvernement ouvert pour renforcer (…) l'accès aux matériaux et résultats de la recherche » (p.7).

Ainsi, dans ce plan d'action, le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation est désigné comme l'institution porteuse de l'engagement 18 pour construire un écosystème de la science ouverte dans lequel « la science sera plus cumulative, plus fortement étayée par des données, plus transparente, plus intègre, plus rapide et d'accès plus universel (et qui) induit une démocratisation de l'accès aux savoirs, utile à la recherche, à la formation, à la société » (p.57).

Parmi les actions de la feuille de route, plusieurs se réfèrent directement ou indirectement aux données de la recherche (p.58) :

- La création d'un comité pour la science ouverte (2018),
- L'enrichissement et le développement d'Isidore, plateforme de recherche permettant l'accès aux données numériques des sciences humaines et sociales (2018-2020),
- La communication auprès des communautés scientifiques sur les implications de la loi numérique relatives à l'ouverture des publications et des données (2018 ou 2019),
- La recommandation de l'adoption d'une politique de données ouvertes associées aux articles et le développement des *data papers*, dans le cadre du soutien public aux revues scientifiques (n.d.),
- La généralisation progressive, via un accompagnement, de la mise en place de plans de gestion des données dans les appels à projets de recherche, et l'incitation à une ouverture des données produites par les programmes financés (2019).

Le plan national pour la science ouverte, présenté début juillet 2018, a confirmé cette ambition[2]. L'objectif est que les données produites par la recherche publique soient progressivement structurées en conformité avec les principes FAIR, préservées et, quand cela est possible, ouvertes. Le plan annonce trois mesures :

- L'obligation d'une diffusion ouverte des données issues de programmes financés sur fonds publics,
- La création d'une fonction d'administrateur de données et le réseau associé au sein des établissements,
- La promotion d'une politique de données ouvertes associées aux articles publiés par les chercheurs.

Ces mesures sont traduites en dix actions à mettre progressivement en place, dès 2018, parmi lesquelles se trouvent

- le lancement d'un appel à projet ANR « Flash » à l'instar du programme européen H2020
- la création d'un prix récompensant les équipes et projets exemplaires
- la généralisation des plans de gestion de données dans les appels à projets

[1] Etalab (2018)
[2] MESRI (2018)

- le développement de centres de données thématiques et disciplinaires et d'un service générique d'accueil et de diffusion de données simples
- l'engagement d'un processus de certification des infrastructures de données.

Le développement d'un écosystème de la science ouverte s'inscrit dans un contexte mondial, caractérisé par la technologie numérique, les réseaux organisés en ligne et les outils collaboratifs. Le plan d'action renvoie notamment vers l'initiative d'un bureau international d'appui et de coordination, (*GOFAIR International Support and Coordination Office - GFISCO*) dont « l'objectif est d'ouvrir progressivement les données de la recherche existantes au sein des institutions scientifiques et académiques, dans tous les domaines de la recherche et au-delà des frontières nationales, constituant un tremplin vers la réalisation de l'Open Science Cloud européen » (p.59)[3].

Aussi, la LERU, la Ligue des Universités de Recherche Européennes, vient de publier une feuille de route (*advice paper*) pour accélérer le développement d'une culture de science ouverte au sein des établissements[4]. La LERU plaide pour un changement culturel et propose un cadre général, à partir des huit axes de la Commission Européenne[5]. Ainsi, la Ligue préconise l'adoption d'une politique institutionnelle en conformité avec les principes de l'UE, la mise en place de services et infrastructures, le signalement des données et la diffusion libre des métadonnées, tout cela accompagné par des formations et des coopérations au niveau local, national et international.

Notre étude *Vers une culture de la donnée en SHS* souhaite contribuer à la mise en œuvre de l'écosystème de la science ouverte sur le terrain d'un campus universitaire. L'étude fait suite à des travaux de recherche menés depuis 2013 par le laboratoire GERiiCO[6], en complément plus particulièrement de l'enquête sur les données de la recherche en SHS sur le campus de l'Université de Lille 3[7] en 2015[8]. Conduite sous forme d'entretiens avec quelques dizaines de chercheurs, doctorants, responsables de laboratoires, chefs de projets et ingénieurs en charge de données, l'étude poursuit trois objectifs :

1. (Re)Mettre les enseignants-chercheurs au cœur de la mise en œuvre de l'écosystème de la science ouverte sur le campus, avec leurs besoins, priorités et interrogations ;
2. Identifier des opportunités et verrous pour une politique de données ;
3. Recommander dix actions à mettre en place pour développer la culture de données sur le campus.

Menée comme un audit sur un terrain particulier[9] et dans le domaine des sciences humaines et sociales, l'étude se réfère à d'autres travaux, dont notamment les enquêtes des universités Rennes 2, Strasbourg, Humboldt de Berlin et Wageningen[10], les études menées par la Direction IST du CNRS[11] et d'autres recherches[12]. Ces travaux ont abouti à l'état des lieux détaillé des pratiques et besoins, ils ont encouragé le développement d'une offre de services de prestations de conseil, d'assistance et de

[3] EOSC cf. http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[4] LERU (2018)
[5] Cf. https://ec.europa.eu/research/openscience/index.cfm#
[6] Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication https://geriico-recherche.univ-lille3.fr/
[7] Avant la fusion de l'Université de Lille en janvier 2018, l'Université de Lille 3 désignait le campus Sciences Humaines et Sociales
[8] Prost & Schöpfel (2015)
[9] Campus Pont de Bois (Villeneuve d'Ascq) de l'Université de Lille
[10] Cf. Serres et al. (2017), Rege (2015), Michel & Chekib (2015), Simukovic et al. (2013, 2014), van Zeeland & Ringersma (2017)
[11] Etudes COPIST et autres, cf. http://www.cnrs.fr/dist/publications.html ; cf. aussi Schöpfel et al. (2018)
[12] Par exemple, Reilly et al. (2011), Bauer et al. (2015)

formation autour des entrepôts, des outils et des plateformes. Ils ont également alimenté la recherche sur les pratiques et usages des chercheurs et sur les dispositifs ; ils ont permis d'approfondir la connaissance de la gestion et la compréhension de la nature même des données.

Notre étude a une portée pragmatique: dégager les éléments indispensables pour une politique cohérente de la production, gestion et réutilisation des données de la recherche sur un campus en sciences humaines et sociales, et contribuer ainsi à l'appropriation du concept de la science ouverte par une « mise en culture de la donnée, qui effectue une mise en sens d'usages disséminés et spécialisés de données ouvertes »[13].

Le plan d'action national 2018-2020 constate qu'il reste « encore beaucoup à faire pour que la science ouverte prenne toute sa place dans les pratiques scientifiques » (p.57). Pour réussir, une telle démarche nécessite une connaissance de la réalité du terrain ; elle a besoin du soutien des communautés de chercheurs, d'une coordination de tous les acteurs sur le campus et d'un pilotage institutionnel et scientifique. Cela prendra du temps[14]. Mais c'est un investissement nécessaire pour maintenir l'excellence de la recherche.

Nous tenons ici à exprimer nos remerciements à tous nos interlocuteurs, aux collègues du laboratoire GERiiCO (en particulier Stéphane Chaudiron, Bernard Jacquemin, Eric Kergosien et Hélène Prost), à Cécile Malleret du SCD pour l'accompagnement de l'enquête, à Leslie Hyacinthe du département SID qui a mené une partie des entretiens en tant que vacataire scientifique, et à la MESHS pour le financement et soutien logistique du projet.

[13] Debos (2017)

[14] Cf. van Zeeland & Ringersma (2017) : "Needless to say, the establishment of RDM policies takes time. It is a lengthy process of which defining the guidelines is only an initial step. Once in place, policies need to be integrated into researchers' behaviour, accepted as part of their culture, and continuously reviewed and audited" (p.165)

Constats préalables

« La science du XXI[e] siècle est la science de l'exploration des données »[15]. Collectées et produites dans le cadre des projets scientifiques, leur gestion pose un défi quotidien et inédit aux chercheurs et aux structures de recherche, aux bibliothèques et aux services centraux des universités : comment stocker, communiquer et conserver ces données en toute sécurité, comment les partager avec d'autres chercheurs ? Comment faire le lien avec les publications ? Comment les intégrer dans une politique du libre accès à l'information scientifique et, plus largement, dans un projet de science ouverte ?

Face aux potentiels et risques technologiques, face aussi au développement de la législation et des programmes de recherche, la mise en œuvre de « bonnes pratiques » dépasse largement le travail scientifique individuel et rend nécessaire une réponse politique et institutionnelle. Cependant, « si les données sont au cœur du travail des chercheurs, les chercheurs sont au cœur de toute politique des données (…) Or, aucune politique de description, de stockage, d'archivage, de partage des données ne pourra se mettre en place sans l'adhésion ou l'implication des chercheurs »[16].

Les chercheurs sont les acteurs-clé de toute démarche qui vise l'excellence de la recherche. Comment obtenir leur adhésion ? Comment les inciter à s'impliquer ? Quelle politique mener pour répondre au mieux à leurs besoins ? L'enjeu de notre étude et de fournir quelques éléments de réponse à ces questions. Commençons par quelques constats à partir des deux seules enquêtes sur la gestion des données de la recherche en SHS, en milieu universitaire et en France, à Lille (2015) et à Rennes (2017).

La longue traîne

De quelles données parlons-nous ? L'enquête de 2015 a révélé un large éventail de données sources sur le campus de l'Université de Lille 3. Les corpus de textes sont la source la plus importante (64%), suivis par les enquêtes et entretiens (47%), observations (41%), expériences (36%) et archives (34%). Quant à la typologie des données produites par les chercheurs comme résultat de leur travail, les données textuelles occupent de loin le premier rang (75%), suivies des tableaux (49%), bases de données (37%), visualisations ou modèles multidimensionnels (32%) et divers formats AV (photo, son, vidéo ; 20-25%).

Par rapport au campus de l'Université Rennes 2, Serres et al. (2017) évoquent la complexité et la spécificité des données de la recherche, en particulier dans les SHS : « Qu'il s'agisse de leurs définitions, difficiles et multiples, de leur distinction pas toujours évidente d'avec les publications, de la difficulté à séparer parfois données collectées et données produites, de l'entremêlement des phases du cycle des données, de leur immense variété... : les données de recherche en SHS ne se laissent pas aisément définir et saisir » (p.118).

Dans le cadre du projet *D4Humanities*, nous avons privilégié une triple approche des données de la recherche, conceptuelle, typologique et fonctionnelle, qui met en avant les liens avec la communauté scientifique et avec les fonctions des données pour la recherche[17]. Il est certain que la gestion des données sur un campus SHS est différente de celle d'un grand équipement ou instrument scientifique (observatoire, synchrotron, supercalculateur etc.). Il n'est pas nécessaire, par contre, de caractériser les données en SHS comme « small data » en opposition au concept du Big Data, puisqu'elles partagent avec le Big Data les caractéristiques essentielles, notamment la volumétrie, la

[15] André (2015), p.79
[16] Serres et al. (2017), p.7
[17] Schöpfel et al. (2017)

diversité et le flux ininterrompu (vitesse). Parlons donc plutôt d'une longue traîne de données, d'un grand volume de données très variées, omniprésentes, reparties sur une multitude de disciplines, de domaines, d'équipes et de projets.

Pratiques

La grande diversité de la nature des données et des conditions de leur production trouve son reflet dans les pratiques de gestion. L'enquête de 2015 en a fourni une image détaillée, composée d'une importante variation de pratiques mais aussi, de certaines caractéristiques prépondérantes. Par exemple, le stockage en local est de loin le mode de sauvegarde privilégié, que ce soit sur l'ordinateur privé ou sur leur ordinateur professionnel. 19% des chercheurs stockent « dans le cloud », alors que 8% ont des données sur le serveur d'une autre institution. En réseau, 12% des répondants se tournent vers le serveur de l'université.

Certaines pratiques de stockage et de communication présentent des risques de sécurité. D'autres ont un caractère plus individuel que collectif. Parfois, ces pratiques ont été décrites comme « artisanales » ; il faut prendre ce terme au sens noble, non comme synonyme pour du bricolage ou de l'amateurisme mais pour un travail de qualité, sur mesure, à partir d'un capital d'expérience partagée et sans objectif d'industrialisation. Ce qui fait la différence, souvent, ce n'est pas le savoir-faire scientifique individuel ou collectif mais la disponibilité ou l'absence de dispositifs et services de données adéquats. Ce qui fait la différence aussi, ce sont les ressources financières et humaines pour la gestion des données[18].

La majorité ne partage pas ses données avec d'autres. Et ceux qui le font, partagent d'abord et surtout avec les collègues de l'équipe scientifique (34%). Très peu (<5%) ouvrent leurs données davantage et partagent avec l'institution, d'autres chercheurs ou « tout le monde », dans une démarche *open data* au sens strict du terme.

Motivations

Les résultats de l'enquête de 2015 laissent penser qu'environ 20-25% des répondants ont davantage d'expérience que les autres (« précurseurs »). Un peu moins de 20% se disent prêts et ont l'intention d'adopter ces pratiques à l'avenir (« motivés »). 30% manquent d'information et de connaissance concernant le partage des données de la recherche (« ignorants »). Et seulement 5-10% indiquent qu'ils n'ont pas l'intention de partager leurs données à l'avenir (« réticents »).

Sans doute, il faut interpréter ces chiffres avec prudence. Serres et al. (2017) avertissent d'un écart voire de contradictions flagrantes entre les déclarations, les représentations et les pratiques en matière de partage, avec des incertitudes et réticences à diffuser les données en libre accès.

En plus, les motivations individuelles ne sont qu'une partie des facteurs. D'autres variables paraissent déterminer au moins autant (et probablement davantage) le comportement de partage, comme les pratiques collaboratives, les usages disciplinaires et communautaires, les incitations des agences de financement et les obligations imposées par la loi.

Attentes

A Lille, l'espace d'archivage sécurisé et fiable figure en tête de la liste des besoins, avant la demande d'assistance et de conseil pour la gestion des données en général, pour des aspects plus techniques (normes, métadonnées etc.) et pour des questions juridiques (autour de 50%). Moins nombreux sont les chercheurs qui souhaitent des conseils ou de l'assistance pour la publication ou diffusion des

[18] Cf. l'évaluation de la situation par les directeurs de laboratoire du CNRS (Schöpfel et al. 2018)

données (autour de 40%), encore moins ceux qui demandent de l'aide pour préparer un plan de gestion ou un conseil dans le domaine de l'éthique.

L'enquête de Rennes a partiellement confirmé cette hiérarchie des besoins, en soulignant l'importance des aspects techniques et juridiques qui sont prioritaires chez les enseignants-chercheurs, en comparaison surtout avec la diffusion en libre accès ou la curation[19]. Un autre trait marquant de cette enquête – peut-être plus important - est la diversité des besoins exprimés car elle pose un problème au développement des services de données sur le campus : faut-il privilégier le développement d'une offre de service large, couvrant l'ensemble des demandes de la part des communautés scientifiques, ou faut-il faire un choix, définir des priorités et miser sur la profondeur de l'offre de service sur certains créneaux mis en avant ?

Services de données

Dernier constat préliminaire, concernant l'offre de service. Les enquêtes ont tendance à focaliser quelques services prioritaires, comme les plateformes de stockage ou les entrepôts d'archivage et de publication[20]. Or, la gamme des services dédiés aux données scientifiques est bien plus large et couvre l'ensemble des phases du cycle de vie des données de la recherche. Le wiki Cat-OPIDoR, un catalogue qui vise à recenser les services français dans ce domaine[21] propose un référencement de neuf catégories de service :

- Sites d'information
- Services de formation
- Aide personnalisée (accompagnement)
- Outils de gestion (plans de gestion de données, éditorialisation, attribution d'identifiants pérennes etc.)
- Plateformes d'acquisition (collecte de données)
- Plateformes de calcul (ressources informatiques pour l'analyse, de la simulation, de la modélisation et du stockage)
- Entrepôts de données (plateformes de dépôt et de partage)
- Annuaires de données (outils de référencement)
- Plateformes d'archivage (conservation des données à long terme)

On peut aller plus loin dans la cartographie de l'offre de services : identifier par exemple les publics cibles (disciplines, rattachement…), le périmètre (national, local…), le statut (mise en œuvre par un SCD, infrastructure nationale, service d'un organisme de recherche…) ou le modèle (offre spécifique, intégration dans un ensemble de service plus large…).

Il ne manque pas de publications, modèles et retours d'expériences sur les services de données.[22] Pour un campus universitaire, le modèle le plus pragmatique et opérationnel suggère une offre de service organisé sur trois niveaux qui s'appuie sur la bibliothèque universitaire[23] :

- Formation (séminaires, journées d'études, stages, ressources en ligne, guides, Q/R, veille…)
- Conseil, assistance (conseil juridique et technique, gestion, DMP, assistance technique pour dépôt, liaison avec laboratoires…)

[19] Serres et al. (2017), p.120-121
[20] *Data repositories* au sens du répertoire international *re3data* cf. https://www.re3data.org
[21] Financé par le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation et hébergé par l'INIST (CNRS), cf. https://cat.opidor.fr
[22] Cf. par exemple Pryor et al. (2014)
[23] *Three tiers model of research data support services*, cf. Reznik-Zellen et al. (2012)

- Infrastructures (médiation avec dépôts de données, partenariats avec réseaux et prestataires, développement d'outils sur le campus…)

C'est ce dernier modèle que nous avons adopté sur le campus de Pont de Bois pour le développement d'une offre de service en destination des doctorants[24].

Pour revenir à l'objectif de notre étude, la gestion des données est un défi quotidien qui mobilise les ressources et dispositifs locaux, mais également les services, outils et infrastructures de niveau national ou international. Les enquêtes et retour d'expériences concordent sur un point essentiel : dans le domaine des données de la recherche, il n'y a pas de solution unique, et une stratégie *top-down* est vouée à l'échec. Pour répondre aux besoins des chercheurs, il faut prendre en compte leurs pratiques. Certaines communautés sont plus avancées que d'autres dans l'archivage et le partage des résultats scientifiques. Dès le départ, il faut abandonner l'idée d'une solution unique en faveur d'une approche modulaire, différenciée, par option.

Pour développer une offre de service utile aux chercheurs, nous devons partir du terrain, évaluer les attentes et usages, intégrer aussi d'autres paramètres, l'existence de réseaux scientifiques, de projets structurants, d'infrastructures disciplinaires. La démarche nécessite la concertation étroite des services concernés (laboratoires, école doctorale, DSI, DGS, DR, SCD) et un pilotage scientifique ; elle mobilisera des partenariats avec des organismes et réseaux opérationnels proposant des ressources et outils pertinents (Huma-Num, DARIAH, CNRS…).

Notre étude n'a qu'une seule et unique ambition : contribuer à améliorer les pratiques sur le terrain et développer une offre de service qui répond aux besoins des chercheurs et aux exigences nouvelles de la politique de la science ouverte. Développer une culture de la donnée, c'est le seul moyen pour faire partie de l'écosystème de la science ouverte, condition *sine qua non* d'une recherche d'excellence du 21e siècle.

[24] Livre blanc sur les données de la recherche dans les thèses de doctorat, cf. Chaudiron et al. (2015)

Eléments méthodologiques

L'objectif général du projet *D4Humanities* est d'accélérer la démarche des données de la recherche de l'Université de Lille dans le domaine des sciences humaines et sociales, notamment par rapport aux doctorants et jeunes chercheurs, et de faciliter le montage d'un projet de recherche international[25]. L'un des volets du projet porte sur les pratiques et besoins dans le domaine des données de la recherche.

Après notre première enquête de 2015, l'objectif est de mener une deuxième étude complémentaire et qualitative, avec un échantillon représentatif mais plus restreint de chercheurs sur le campus de Pont de Bois. Au départ, l'intention est double :

1. D'une manière générale, nous souhaitons approfondir et enrichir la connaissance des comportements, attitudes, motivations et besoins par rapport à la gestion et au partage des données de la recherche.
2. En particulier, nous voulons évaluer les attentes vis-à-vis de l'offre de services à la recherche, et nous souhaitons appréhender comment est comprise l'application des principes de la gestion des données dans l'environnement émergent de la Science ouverte (*FAIR data principles*[26]).

La trame pour des entretiens semi-directifs d'une durée d'environ une heure a été préparée au cours du premier semestre 2017 (Annexe 1). Les 40 questions de la trame abordent huit thématiques (entre parenthèses, le nombre de questions) :

1. Informations sur la personne interrogée (5)
2. Implication dans des projets de recherche et connaissance des consignes en matière de données (4)
3. Nature des données collectées et générées (5)
4. Expérience des plans de gestion et d'autres outils (5)
5. Description et organisation des données (4)
6. Partage et diffusion des données (9)
7. Stockage des données (4)
8. Besoins en matière de données (4)

La trame de l'entretien a été testée auprès de quelques volontaires; après une première série d'entretiens, plusieurs questions ont été reformulées et enrichies. Les entretiens ont été menés par l'auteur de l'étude et une vacataire de recherche (Leslie Hyacinthe). La plupart des entretiens ont été retranscrits via un outil d'enquête en ligne. Cependant, certains entretiens ont été conduits sous forme d'audit, suivant une partie seulement des questions, en ajoutant d'autres questions en fonction de l'expertise des personnes interrogées.

L'établissement de l'échantillon a suivi plusieurs logiques :

- Appel aux répondants volontaires de l'enquête 2015
- Appel aux doctorants volontaires du séminaire doctoral DRTD de 2016-2017
- Appel aux participants de plusieurs événements de formation du SCD (Pause-Doc et autres)

[25] https://d4h.meshs.fr/
[26] "Findable, Accessible, Interoperable and Reusable" cf. Wilkinson et al. (2016)

- Constitution d'un échantillon aléatoire à partir des annuaires des laboratoires du campus Pont de Bois
- Contact direct auprès des porteurs de projets scientifiques nationaux (ANR) et européens
- Contact direct de plusieurs directeurs de laboratoires
- Contact direct avec plusieurs experts en matière de données, notamment de la Direction de la Recherche et de la Direction des Systèmes d'Information

En fin de compte, nous avons mené 51 entretiens d'une durée variable, entre une et deux heures, dont 43 entretiens individuels et quatre entretiens avec deux interlocuteurs. L'impact de la fusion des trois universités de Lille n'a pas facilité la disponibilité des uns et des autres. Les entretiens ont été organisés en trois phases, au printemps 2017, durant le deuxième semestre 2017 et début 2018. La liste des interlocuteurs se trouve en annexe 2.

Les personnes interrogées proviennent de douze équipes de recherche ou laboratoires différents ; les quatre premiers laboratoires où nous avons rencontré le plus de répondants sont respectivement SCALab (Sciences affectives et cognitives), GERIICO (Groupement d'études et de la recherche interdisciplinaire en information et communication), CECILLE (Centre d'études en civilisation, langues et lettres étrangères), et HALMA (Histoire, archéologie et littérature des mondes anciens). Par contre, les services centraux du campus font partie de notre panel, représentés par la Direction Générale des Services, la Direction de la Recherche et la Direction des Services Informatiques. Quant au statut des répondants, la plupart sont enseignants-chercheurs (43%), doctorants (20%) ou directeurs de laboratoire (18%). Un tiers des personnes interrogées ont une responsabilité dans la direction d'une structure de recherche ou de mission.

Observations : un paysage contrasté

A partir des constats préalables, nous avons mené 51 interviews pour mieux comprendre la réalité de la gestion des données de la recherche sur le terrain des sciences humaines et sociales. Nous allons résumer le résultat de ces entretiens sous forme de huit observations. Cependant, les difficultés d'organisation de ces entretiens ont fourni le premier résultat de notre étude.

En effet, dans un premier temps nous avions pris contact avec 72 enseignants-chercheurs à partir d'un échantillonnage aléatoire, basé sur les listes des membres des laboratoires du campus de Pont de Bois. Malgré deux relances, cette prise de contact fut un échec. Trois destinataires sur quatre n'ont tout simplement pas répondu, et une partie des autres ont fini par répondre qu'ils n'avaient pas le temps de participer à cette étude.

On peut interpréter cette réaction de plusieurs façons. La préparation de la fusion des trois universités lilloises n'a sans doute pas favorisé la disponibilité et la motivation des professeurs et maîtres de conférences contactés ; la période de la rentrée non plus. Et on peut toujours supposer que l'information sur l'enquête n'a pas été assez complète et convaincante, que la communication n'a pas été optimale, etc.

Néanmoins, d'autres échanges en amont et durant les entretiens laissent penser qu'il y a (aussi) une autre raison : en fait la gestion des données de la recherche n'est pas considérée comme une priorité absolue mais plutôt comme un sujet subi, une actualité imposée par les programmes de recherche européens, par la réglementation européenne et nationale autour des données personnelles, et par des partenaires scientifiques, dont notamment le CNRS et les hôpitaux.

Aussi, plusieurs répondants ont explicitement indiqué qu'ils étaient disponibles surtout par curiosité et par l'opportunité offerte par l'interview de s'informer sur un sujet mal maîtrisé et mal connu, et non pas par la motivation d'échanger sur leurs propres pratiques.

(1) La sécurité des données et des systèmes

Chaque entretien avec les enseignants-chercheurs le confirme : quand il s'agit de leurs données, la première préoccupation n'est pas la gestion en tant que telle, la conservation ou le partage, mais la sécurité des données et, dans un sens plus large, la sécurité des dispositifs utilisés pour leur stockage et leur analyse. Cette observation rejoint la conclusion de l'enquête de l'Université Rennes 2 : « Le manque de sécurité des données est (…) sans nul doute l'un des points les plus cruciaux mis en exergue : il y a réellement urgence à donner une véritable sécurité au stockage et à l'archivage de nombreux jeux de données en SHS »[27].

Aussi, le premier et l'unique interlocuteur en matière de gestion de données est souvent le responsable de la sécurité des systèmes d'information (RSSI) qui garantit la sécurité, la disponibilité et l'intégrité du système d'information et des données sur le campus.

L'urgence de trouver une solution pour sécuriser le stockage et le traitement des données est régulièrement liée aux contraintes légales pour protéger les données personnelles, et en particulier aux obligations de mettre en place quelques précautions et actions élémentaires pour garantir un niveau de sécurité minimal des données personnelles, selon les termes de la Commission Nationale de l'Informatique et des Libertés (CNIL). Plus de la moitié des enseignants-chercheurs de notre panel travaillent avec des données personnelles, que ce soit des données médicales, des photos ou vidéos, ou des comptes Facebook. Du coup, le correspondant informatique et liberté (CIL) est sollicité dans

[27] Serres et al. (2017), p.118

sa fonction d'accompagnement et de conseil, une sollicitation qui correspond tout à fait à ses compétences mais qui dépasse souvent sa disponibilité et ses ressources réelles.

En matière de sécurité, quelles sont les questions abordées avec le RSSI et le CIL ? D'après nos interlocuteurs, il s'agit de plusieurs enjeux :

- Le choix des dispositifs, en particulier quand il s'agit d'une solution « in the cloud ».
- La protection des fichiers, y compris par le cryptage.
- Le stockage sur plusieurs dispositifs et les sauvegardes.
- L'authentification des utilisateurs et la gestion des accès.
- L'anonymisation des données personnelles.

Dans certains cas, il est facile de trouver une solution. D'autres situations posent davantage de problèmes, surtout quand il s'agit de gros volumes de données, d'applications associées et d'un stockage ou traitement sur plusieurs sites. Parfois aussi, l'avis du RSSI est sollicité seulement après un incident, une perte de données, un vol d'équipement etc. ; son rôle se limite dans ce cas au conseil pour réduire le risque d'un nouvel incident.

Cependant, le problème de la sécurité n'est pas limité aux données personnelles ou aux projets de recherche. Ainsi, nos interlocuteurs ont également évoqué d'autres situations nécessitant des solutions sécurisées, comme par exemple, les sites web et serveurs de revues de laboratoire, les bases de données produites au sein d'un laboratoire d'une manière récurrente ou dans le cadre d'un projet ANR, les corpus mis en ligne sous forme de bibliothèque numérique (avec le logiciel Omeka ou autres), les dictionnaires, traductions etc.

Pour l'enseignant-chercheur du terrain et a fortiori pour les responsables d'équipe et de laboratoire, ces problèmes forment un tout et constituent une préoccupation permanente. Selon les termes d'une collègue : *Trouver une solution pour la préservation est une chose, de préférence une solution « clé en main » ou avec des parcours balisés, une offre institutionnelle avec une partie personnalisable ; mais l'autre grand problème, c'est la construction, production et maintenance des bases de données*. La sécurité des données de la recherche n'est pas considérée à part des autres problèmes mais elle risque, par le biais des nouvelles exigences et contraintes, d'exacerber l'impression d'un manque de sécurité informatique sur le campus.

(2) Une communication compliquée

La deuxième préoccupation des chercheurs est la communication des « données chaudes » (en cours de traitement) tout au long du projet, au sein de l'équipe scientifique. Il ne s'agit pas d'un partage de données au sens d'une ouverture ou d'une publication vers un public plus large, mais d'un échange ou transfert de données dans le cadre du cahier des charges d'un projet de recherche.

La question paraît simple au premier abord : comment communiquer des résultats d'une manière efficace au sein d'un projet ? En fait, elle est bien plus compliquée car en fonction de l'envergure, du domaine et de la nature des données du projet, d'autres facteurs entrent en jeu, comme :

- La protection des données personnelles.
- Le consentement des personnes impliquées.
- Le statut des partenaires.
- La localisation géographique et la nationalité des partenaires.
- La sécurité des modes de communication.
- Le risque de perte, de vol, de piratage des données.

Du point de vue « données », la question cruciale est bien l'opposition entre « besoin de communication » et « protection imposée », que ce soit avec les partenaires d'un projet ou à destination d'un public plus large, et cette question se pose pendant tout le processus de la recherche, pas seulement à la fin d'un projet (« données froides »). Dans nos entretiens, les problèmes de communication reviennent comme un leitmotiv, juste après l'enjeu de sécurité.

- Comment communiquer au sein d'un projet où sont impliqués des chercheurs sur plusieurs sites ?
- Comment communiquer au sein d'un réseau ?
- Comment communiquer dans un projet public-privé avec des données produites par ou avec un acteur du secteur privé ?
- Comment communiquer avec des partenaires étrangers, au sein de l'Union Européenne ou dans d'autres pays ?

En absence d'outils de travail adaptés, les équipes ont recours à des solutions commerciales, gratuites ou pas, sur des plateformes. Parmi les outils mentionnés, notamment pour des documents ou des fichiers d'un certain volume, figurent Dropbox et Google Drive. Parfois, une équipe fait une distinction claire entre documents (articles etc.) à communiquer, via Dropbox par exemple, et des données brutes ou codées, sécurisées sur l'ordinateur personnel d'un chercheur. D'autres modes de communication sont l'échange de la clé USB ou du disque externe, ou tout simplement la messagerie électronique.

Ces choix ne s'opèrent pas par naïveté mais en absence d'une alternative efficace et adaptée. Les lignes rouges sont atteintes quand l'un des partenaires s'interdit l'usage de ce genre d'outil, quand il s'agit de données sensibles (caractère personnel, santé, enfants mineurs etc.) ou quand les partenaires se trouvent dans plusieurs pays, avec des cultures et contraintes différentes.

Ces questions se posent par exemple dans les projets du Programme de Coopération Transfrontalière (Interreg France-Wallonie-Vlaanderen) financé par le Fonds européen de développement régional (FEDER). Mais elles se posent avant tout dans les grands projets internationaux regroupant des établissements de plusieurs pays européens et non-européens, mais aussi dans des projets liés à l'industrie ou au domaine biomédical.

(3) Un continuum de pratiques

Comment décrire la réalité des pratiques de gestion des données de la recherche ? Les enquêtes ont tendance à produire une représentation abstraite des pratiques, à partir de valeurs moyennes et d'actes isolés, sorti de leur contexte et sans interconnexion. Une telle approche permet d'identifier et d'observer certains types de pratiques comme sous un microscope, pour mettre en place des prestations et dispositifs ciblés.

Si on met l'accent sur les interconnexions des pratiques, on peut décrire des profils types, des « jeux de pratiques » cohérents qui permettent de distinguer plusieurs groupes de chercheurs (des « clusters »), ce qui peut s'avérer utile pour mieux cibler et ajuster une offre de service modulaire.

On pourrait aussi essayer d'établir un référentiel d'activités et de compétences liées à la gestion des données de la recherche, à la manière des référentiels de métiers ou d'emplois-types, pour obtenir une cartographie fine et exhaustive du domaine de la culture de la donnée, par discipline, équipement, objectif ou métier, ce qui pourrait aboutir à une sorte d'anthropologie des données de la recherche.

Une telle approche pourrait par exemple être utile pour décrire d'une manière globale et fonctionnelle toute la panoplie de pratiques de stockage, de conservation et d'archivage de données, en fonction des critères suivants : le type de données, les injonctions (partenaires industriels, agences de financement) ou obligations (législation), les dispositifs et ressources informatiques (outils personnels ou institutionnels, outil local ou outil « in the cloud » etc.), la volumétrie et la finalité par rapport aux processus de la recherche (stockage temporaire pendant la durée d'un projet ; curation, conservation et exploitation des métadonnées dans un système de recherche ou dans un entrepôt de données, etc.). Mais souvent, comme nous l'avons déjà mentionné, il manque une solution pour un archivage à long terme, avec gestion des accès et usages.

L'approche qualitative de nos entretiens dessine un modèle multidimensionnel de la gestion des données de la recherche, autour de plusieurs axes ou continuums de pratiques, dont notamment :

Formalisation : certaines pratiques relèvent davantage d'une construction personnelle et sur mesure, tandis que d'autres ont un caractère plus formalisé, comme par exemple la compilation de données de géolocalisation pour alimenter une carte archéologique, ce qui nécessite une procédure collective et acceptée. Dans certains cas, les chercheurs disposent de guides, de manuels ou de recommandations, comme pour l'analyse statistique, pour le traitement des données issues d'entretiens ou encore dans le domaine de la santé. Une minorité d'enseignants-chercheurs attribuent des codes pour identifier leurs données, afin de garantir l'anonymat, selon des règles d'usage. Pour décrire les données, cinq éléments prédominent : le nom de fichier, l'auteur, le nom du projet (avec financement et date), le domaine thématique, les droits d'usage (protection, confidentialité).

Finalité : on retrouve ici le lien avec le cycle de vie des données, dans la mesure où certaines pratiques correspondent au traitement (préparation, nettoyage, analyse, interprétation) des données tandis que d'autres appartiennent davantage à la gestion des données : par exemple la description des droits d'accès et d'usage des enregistrements audio-visuels d'entretiens ou d'interventions dans le domaine de la santé (à partir des consentements obtenus). Un autre exemple est l'indexation ou le codage des contenus d'une vidéo en appliquant les recommandations du domaine (types d'intervention etc.).

Adéquation : certaines pratiques sont vécues comme peu satisfaisantes (le stockage sur disque dur de l'ordinateur personnel, la communication des données via Google Drive ou Dropbox etc.) tandis que d'autres pratiques sont considérées en adéquation avec les besoins et contraintes (l'anonymisation des données personnelles avec un algorithme par exemple, ou la génération des noms de fichiers, la diffusion via le cloud de l'Université de Lille ou de Valenciennes, le partage sur GitHub). On pourrait citer ici aussi l'utilisation d'outils comme le *MIT Saliency Benchmark*, avec des données de référence dans le domaine des expériences physiologiques (suivi de l'œil). Un autre vecteur de partage est peu évoqué mais paraît en adéquation avec les besoins des chercheurs : les réseaux sociaux de recherche, notamment Academia, proposant une diffusion à la demande.

Quand les enseignants-chercheurs évoquent des consignes pour les données, il s'agit de trois domaines : les données médicales, les revues en psychologie, les règles pour décrire les corpus linguistiques et textuels. Quant aux plans de gestion, environ un tiers des personnes interrogées indiquent avoir rédigé un tel document, sous une forme ou une autre.

(4) Plusieurs niveaux de gouvernance

Qui est responsable des données de la recherche ? Qui coordonne leur gestion ? Qui a une vue globale de la collecte et production des données ? Les réponses confirment globalement le constat

de l'enquête de 2015 : les données sont souvent considérées comme une affaire personnelle, voire comme une propriété privée, sous la responsabilité du chercheur. Mais à travers les réponses se dessine aussi une autre réalité, celle d'une gouvernance à plusieurs niveaux – une gouvernance parfois structurée et explicite, mais souvent diffuse et plus ou moins informelle.

D'une manière générale, on peut distinguer deux niveaux, le projet et le laboratoire. Cependant, la réalité est plus complexe. Au sein d'un projet, les entretiens révèlent trois situations types :

- Par défaut, la responsabilité globale des données est attribuée au directeur du projet, à la personne qui a coordonné le montage du projet, qui a une vision globale du projet et qui est responsable vis-à-vis de l'agence de financement. C'est elle ou lui qui est censé(e) savoir comment gérer les données, comment sécuriser leur stockage et communication au sein de l'équipe projet, comment garantir leur préservation à la fin du projet etc.
- D'autres projets ont désigné une personne responsable des données parmi les membres du projet. Il peut s'agir d'un chercheur, d'un informaticien ou d'un ingénieur de données (SHS, documentation). Dans les projets européens, cette désignation est formalisée par le plan de gestion. Dans d'autres projets, le protocole éthique remplit ce rôle.
- D'autres projets encore n'ont ni « leader » ni « data officer » mais semblent gérer leurs données au fil de l'eau, selon un partage entre responsabilité individuelle (chacun est son propre « data officer ») et coordination élémentaire au sein de l'équipe projet.

La qualité de la gouvernance et le degré d'une structuration dépend du programme de recherche, de l'envergure du projet et de la nature des données (personnelles, biomédicales, sensibles…). L'effet structurant des projets, notamment européens, sur la gouvernance des données, paraît important.

Le laboratoire ne joue pas le même rôle. Rares sont les laboratoires déclarant une politique ou gouvernance explicite en matière de données. L'alimentation centralisée et coordonnée d'une base de données regroupant les résultats des travaux de recherche des membres d'un laboratoire, comme en archéologie, reste l'exception. La tendance est plutôt d'attribuer la responsabilité des données aux chefs de projet et aux chercheurs individuels.

Ceci étant, les directeurs de laboratoires peuvent être un vecteur de communication pertinent quand il s'agit de sensibiliser les chercheurs aux enjeux de la gestion des données de la recherche. Ainsi, surtout quand ils forment une sorte de comité de coordination permanent comme en section 71[28] ils peuvent fonctionner comme un « levier politique » au sein d'une communauté disciplinaire.

Au moment de l'enquête, il n'y avait pas de gouvernance de donnée au niveau de l'établissement (Université de Lille 3) ou du campus (Pont de Bois). Entre temps (février 2018), la nouvelle Université de Lille a nommé un chargé de mission pour les données de la recherche, ce qui peut signifier le début d'une gouvernance, y compris pour les données en SHS.

(5) Disciplines ou méthodes ?

La gestion des données de la recherche dépend pour beaucoup de la discipline scientifique : cette affirmation fait partie des lieux communs de la littérature sur les données de la recherche. Ce constat n'est pas faux ; et il est certain qu'on gère les données différemment en histoire et en physique nucléaire. Aussi, l'expression des besoins de conseil se fait souvent dans un cadre disciplinaire – des logiciels pour tel domaine, des lieux de stockage pour telle disciplines etc.

[28] Conférence Permanente des Directeurs.trices de laboratoires en Sciences de l'Information et de la Communication cf. http://cpdirsic.fr/

A regarder de près, et en limitant le périmètre de l'analyse aux seules sciences humaines et sociales, il apparaît néanmoins que quand bien même la discipline constitue l'une des variables indépendantes de la pratique sur le terrain, deux autres facteurs exercent une influence déterminante (figure 1) :

- D'une part, les méthodes et équipements : les enquêtes qualitatives, les fouilles archéologiques, les protocoles d'expérimentation neuropsychologique, les analyses de réseaux etc.
- D'autre part la réglementation pour certains types de données (données personnelles), personnes (mineurs) ou traitements.

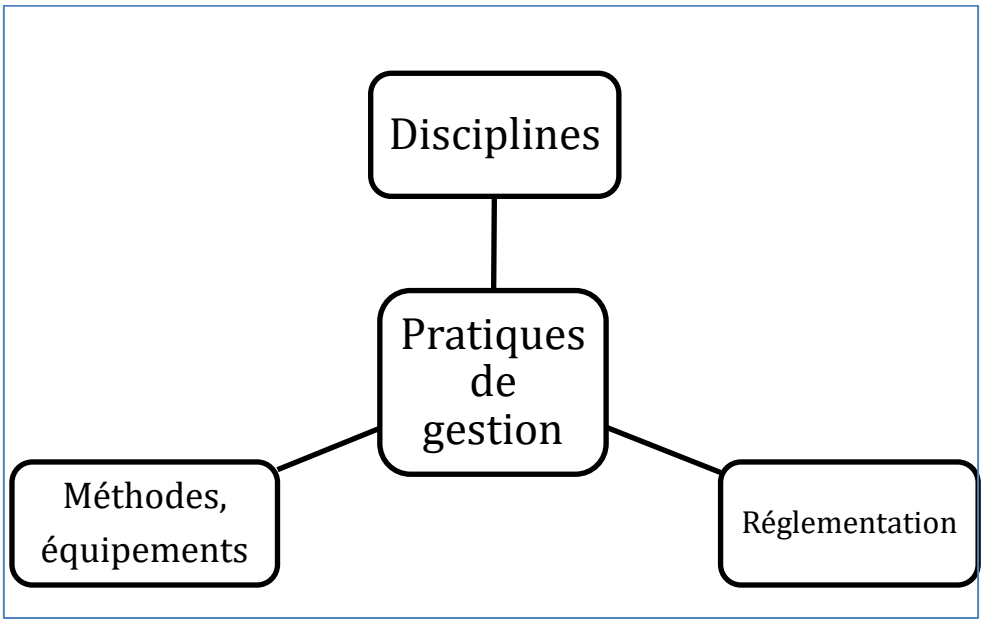

Figure 1. Facteurs indépendantes des pratiques de gestion.

Ces deux facteurs sont certes liés aux thématiques et domaines d'investigation, mais ont aussi un caractère fondamentalement transdisciplinaire. Concrètement et par rapport aux entretiens menés dans notre étude, cela signifie que les chercheurs de la même discipline ou du même laboratoire n'ont pas nécessairement les mêmes pratiques et approches des données ; et cela veut dire aussi que d'autres chercheurs appliquent les mêmes « stratégies de données » du fait d'une même approche méthodologique (par exemple, des enquêtes, des enregistrements d'entretiens, ou des analyses de réseaux) et ceci malgré une origine disciplinaire différente.

(6) Questions terminologiques

Pour communiquer, il faut se comprendre. Un minimum de terminologie partagée et de cadre de référence commun est d'autant plus indispensable quand il s'agit d'informer et d'évaluer. Or, dans le domaine de la gestion des données de la recherche, ce n'est pas une évidence. Le cadre théorique, les thématiques et les méthodes de nos interlocuteurs sont très variés : citons au hasard : : des statistiques spatiales, des données EMG et EEG, les processus de jugement, les songes dans la littérature, la métallurgie préhistorique, Victor Hugo, le transfert des connaissances chez des jeunes autistes, les discours médiatiques, la maladie dAlzheimer, des réunions de groupe focus, des photos de structures architecturales, la scientométrie, l'empathie des médecins, la culture de l'information ou encore les temples égyptiens. Trouver une base d'entente partagée dans le domaine de la donnée n'est pas une évidence.

Dans nos entretiens, nous avons régulièrement constaté une incompréhension de deux concepts, celui de la donnée et celui du partage.

Concept de la donnée

Comme dans les journées d'étude, une des premières questions dans un entretien sur la gestion des données a été souvent : « Et pour commencer, pourriez-vous définir 'données de la recherche' ? » (cf. Schöpfel et al. 2017). Pour une partie des chercheurs et doctorants interrogés, ce terme de « donnée de la recherche » ne semble pas avoir beaucoup de sens car trop abstrait, trop large et inclusif, sans lien avec les « objets » qu'ils utilisent, analysent et produisent.

Quand ils parlent de données, comme en psychologie ou en archéologie, c'est par rapport aux bases de données, parfois aussi, comme en littérature, par rapport aux références bibliographiques. Pour le reste, la terminologie est celle des objets manipulés, comme des cartes, transcriptions, enregistrements physiologiques, statistiques, observations, textes etc.

De même, la distinction entre données « chaudes » et « froides » n'est pas abordé dans ces termes. Pourtant, la différence est bien connue entre un stockage temporaire pour la durée du projet, avec des mises à jour, des traitements, des versions successives etc., et l'archivage définitif une fois le projet terminé et les résultats publiés. Mais les questions se posent d'une manière concrète, par rapport à des données réelles et dans l'environnement d'un projet particulier, pas d'une manière générale.

Ce manque partiel de cadre de référence commun concerne également les plans de gestion. La plupart de nos interlocuteurs n'ont jamais rédigé un plan de gestion et se demandent à quoi cela peut bien servir. Or, un certain nombre, notamment en psychologie, a déjà dû préparer un protocole éthique à destination du comité d'éthique, et une partie de ce protocole correspond à un plan de gestion des données, avec description des données, de la méthodologie de leur production et collecte, de leur traitement, de leur stockage et archivage etc.

Concept du partage

Nous l'avons indiqué plus haut : peu de chercheurs sont réfractaires à l'idée de partager leurs résultats scientifiques ; en même temps, déposer des données dans un entrepôt pour une diffusion libre soulève des interrogations et reste une exception.

L'équipe de Rennes 2 évoque ici un écart voire des contradictions entre les déclarations, les représentations et les pratiques des chercheurs[29]. Nos entretiens montrent un paysage plus contrasté. Contrairement à l'idée reçue d'une dichotomie entre une science « open » et une science « closed », la plupart des chercheurs communiquent leurs résultats, sous forme de publications, mais aussi sous d'autres formes – bases de données, tableurs, listes, collections etc. Pour eux, le partage ne se pose pas en termes de « communication oui/non » mais d'une façon différenciée, suivant trois questions :

- Communiquer avec qui ?
- Communiquer quoi ?
- Communiquer de quelle manière ?

Aussi, le partage est considéré comme une forme de travail collaboratif et de communication des résultats entre collègues et partenaires, dans la mesure du techniquement et légalement possible, souvent avec les collègues du même projet et/ou de la même équipe, parfois à la demande.

[29] Cf. Serres et al. 2017

Les finalités sont d'abord le travail scientifique, puis l'évaluation. On partage des résultats pour pouvoir travailler ensemble sur les mêmes jeux de données et aussi pour avoir l'avis des collègues sur la qualité des données et des traitements. Il ne s'agit généralement pas d'un « don symbolique », politique, humaniste ou par conviction éthique, mais d'une pratique intéressée, basée sur la réciprocité. Aussi, peu d'interlocuteurs ont déjà abordé la problématique des licences ouvertes (Creative Commons, GNU).

Les chercheurs évoquent plusieurs conditions et limites pour le partage de leurs données, dont l'autorisation (consentement) et l'anonymisation en cas de données personnelles (y compris au sein de la même équipe), la non-confidentialité des données, mais aussi le « contrôle du public », c'est-à-dire la limitation de l'accessibilité aux collègues scientifiques du même domaine ou du même établissement (ou campus), ou bien après une autorisation du cas par cas, sur demande individuelle. Plusieurs chercheurs disent clairement qu'ils se sentent « responsables de leurs résultats » et qu'ils n'allaient pas « donner les résultats à n'importe qui », en insistant sur le caractère et contexte scientifique des données. Ainsi, une base de données peut se trouver sur un serveur de laboratoire, avec un accès limité à des utilisateurs identifiés, enregistrés et autorisés.

Ces réponses révèlent un autre aspect des données, décrit par Borgman (2016) : la conscience de la valeur réelle et/ou potentielle des résultats de la recherche (« asset » chez Borgman), pour les collègues de la même équipe, pour d'autres chercheurs spécialistes du même domaine et dans certains cas aussi pour un public plus large (santé publique). Moins souvent est évoquée la valeur pour le ou la scientifique à titre individuel – pour « marquer le terrain », pour publier, dans un cas aussi pour construire un ensemble de jeux de données et capitaliser (« thésauriser »), au bout d'une dizaine d'années, sur les résultats de plusieurs projets de recherche. La crainte est le risque d'une appropriation abusive des résultats par quelqu'un d'autre.

(7) Incitations

Quels facteurs favorisent les bonnes pratiques en matière de données, au sens d'une gestion compatible avec les principes FAIR ? Ou d'une manière plus large, quelles sont les raisons qui incitent les chercheurs à mettre en œuvre une gestion réfléchie de leurs données scientifiques ? D'après nos réponses, il s'agit de six facteurs qui peuvent se superposer :

1. **Le programme H2020.** La Commission Européenne a rendu obligatoire, dès 2016, l'accès ouvert aux résultats des projets du programme H2020 et demande un plan de gestion des données évolutif, à trois moments du cycle de la recherche[30]. La Commission avance quatre raisons pour cette ouverture par défaut : la réutilisation des données antérieures pour augmenter la qualité des résultats, l'incitation à la collaboration pour une plus grande efficacité scientifique, une innovation facilitée pour contribuer à la croissance économique, et le développement de la science citoyenne avec une transparence renforcée[31]. Par ailleurs, l'obligation d'un plan de gestion de mi-parcours exprime la volonté d'un suivi qui intègre l'aspect éthique (cf. plus loin). Cette politique porte ses fruits (DCC 2018). Pour le moment, peu de projets sont concernés sur le campus SHS. Parmi nos interlocuteurs, un sur quatre ou cinq seulement ont déjà participé à un tel projet. Mais tous les chercheurs et ingénieurs impliqués dans la préparation et le montage d'un tel projet connaissent cette condition, et ceux qui participent à un projet H2020 sont obligés d'adapter leur pratique à la politique de

[30] « Open access to research data (…) becomes applicable by default in Horizon 2020 » (European Commission 2016, p.3)

[31] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

données de la Commission, même si la responsabilité de la gestion peut se trouver ailleurs, prise en charge par d'autres partenaires scientifiques.

2. **Projets ANR.** La majorité des enseignants-chercheurs ont déjà participé au montage et, moins, à la gestion d'un projet financé par l'ANR. Pour certains, des montages ont été infructueux et/ou remontent à plusieurs années. L'ANR incite les responsables de projet au dépôt d'un plan de gestion des données, avec une ouverture des données à la clé. Mais cela reste une incitation bien moins forte que celle du programme H2020.
3. **Fonctionnement de projet.** Une autre raison pour une gestion réfléchie est liée au fonctionnement et au management d'un projet de recherche, en particulier des projets d'envergure avec plusieurs partenaires et sur plusieurs sites. Comment partager les données ? Comment échanger les fichiers ? Où stocker les jeux de données ? Parfois aussi certains partenaires ont des exigences particulières par rapport à la gestion des données, notamment dans les projets impliquant des partenaires privés et/ou industriels.
4. **La nature des données.** La réglementation pour certains types de données impose une gestion particulière, normalisée, contrôlée. Il s'agit avant tout de deux types de données, les données personnelles et les données liées à la santé. Ainsi, le Code de la Santé Publique (CSP) et la Loi Jardet de 2012 imposent certaines règles quand la recherche implique « la personne humaine ». Des recherches portant sur la sensibilisation au handicap ou évaluant l'efficacité des interventions auprès d'enfants ou adolescents autistes soulèvent des questions précises dans le domaine de la santé et impliquent les Comités de Protection des Personnes (CPP) : ils sont chargés d'émettre un avis préalable sur les conditions de validité de toute recherche impliquant la personne humaine, au regard des critères définis par l'article L 1123-7 du CSP. Les termes de « recherche impliquant la personne humaine » désignent, tout essai ou expérimentation organisé et pratiqué sur l'être humain, en vue du développement des connaissances biologiques ou médicales. D'autres règles s'appliquent à des recherches n'impliquant pas la personne humaine[32]. Par ailleurs, les comités ont la possibilité de faire des audits ponctuels, ce qui renforce l'incitation des bonnes pratiques de gestion.
5. **Protocole éthique.** La nécessité de rédiger un protocole éthique à destination du comité d'éthique de l'établissement (au moment de l'enquête, le comité de l'Université de Lille 3 ; aujourd'hui celui de l'Université de Lille) oblige les responsables de projet à faire le point sur la gestion et le traitement de leurs données, même s'ils ne font pas nécessairement le lien entre les plans de gestion de données et le point 4 du protocole éthique, qui porte explicitement sur le « traitement des données » et sur la sécurité du système d'information. En revanche, ce lien est clairement établi dans le cadre des projets H2020. Ceci étant, l'avis du comité est conditionné par une vérification informatique au préalable (correspondant CNIL, aujourd'hui responsable RGPD), et son caractère est consultatif, sans suivi, à l'exception des projets européens, où l'obligation d'un plan de gestion de mi-parcours comporte un volet éthique. Ceci étant, il faut tenir compte d'autres acteurs et instances auxquels font face des enseignants-chercheurs, comme par exemple le Comité consultatif sur le traitement de l'information en matière de recherche (CCTIRS), l'Inspection de l'Education Nationale en charge des médias, ou le comité d'éthique d'une université partenaire (ici, l'Université du Québec).
6. **Politique éditoriale des revues.** Dernière variable mentionnée par nos interlocuteurs est la demande des comités de rédaction d'informer sur l'accessibilité des données et sur l'avis du comité d'éthique. Dans notre enquête, seul les psychologues ont évoqué cette contrainte,

[32] Comme l'avis du Comité d'Expertise pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé (CEREES), cf. https://www.snds.gouv.fr/SNDS/Actualites/Actu-2

avec deux exemples à la clé : la politique en matière de « data availability » de la revue *PLoS*[33] et les instructions aux auteurs des revues de l'American Psychological Association (APA). Par ailleurs, l'étude sur la politique éditoriale de la revue *Cognition* montre l'impact d'une incitation forte (obligation) sur le partage des données de la recherche, sur le nombre autant que sur la qualité des dépôts[34].

A l'avenir, on peut anticiper une incitation encore plus forte, aussi bien par les instances politiques (Plan d'action 2018-2020) et par la législation (RGPD) que par les agences de financement, les éditeurs et les instances locales qui pourraient, par exemple, encourager le comité d'éthique d'instaurer une sorte de « petit suivi » d'après un calendrier prévisionnel pour le traitement et la conservation des données.

(8) Verrous

Les enquêtes et études sur la gestion des données donnent parfois l'impression que l'obstacle majeur pour une bonne gestion est l'absence de motivation et/ou de compétences des chercheurs eux-mêmes. Les entretiens sur notre campus dessinent une autre image. La réduction psychologique et l'attribution culpabilisante d'une responsabilité individuelle ou collective semble plutôt l'arbre qui cache la forêt, et cette forêt, c'est l'absence de ressources informatiques et humaines sur un campus SHS, en particulier dans les laboratoires universitaires (équipes d'accueil).

1. **Ressources informatiques.** Les interlocuteurs mentionnent en vrac le manque d'espace de stockage en dehors des applications ; l'absence d'outils appropriés pour la communication, le transfert et/ou le partage de jeux de données ; plus en amont aussi des moyens insuffisants pour la construction et la maintenance de bases de données et d'autres applications, avant même d'envisager leur diffusion. La situation paraît plus favorable dans des unités mixtes qui ont recours aux moyens informatiques du CNRS. Cette absence de ressources informatiques est ressentie comme verrou et amène certains chercheurs de se tourner vers des solutions externes, indépendantes, libres (gratuites) ou commerciales.
2. **Ressources humaines.** Plusieurs services et fonctions pénalisés par des ressources humaines insuffisantes pour une bonne gestion ont été identifiés, en vrac :
    - Informatique : sécurité des systèmes d'information, développement et maintenance.
    - Affaires générales : informatique et liberté, aujourd'hui protection des données personnelles.
    - Recherche : montage, suivi et valorisation des projets ; suivi du comité d'éthique.
    - Documentation : administration de bases de données (surtout au sein des laboratoires) ; archivage (conservation).

Les collègues concernés savent généralement ce qu'il faudrait faire pour assurer un minimum de bonne gestion des données de la recherche, mais ils se trouvent souvent dans l'impossibilité de le faire correctement, faute de moyens et confrontés à une charge de travail trop importante. Ce même constat a d'ailleurs été fait par les directeurs d'unités du CNRS lors d'une enquête en 2014[35].

Parmi les problèmes relevés, on trouve l'absence de dispositifs et d'aide pour les doctorants, un nombre insuffisant d'interventions dans les formations universitaires, les problèmes de conservation à long terme sur les serveurs de laboratoire, un manque de suivi des aspects éthiques et

[33] http://journals.plos.org/plosone/s/data-availability
[34] Hardwicke et al. (2018)
[35] Cf. Schöpfel et al. (2018)

informatiques dans les projets d'envergure (audits ponctuels, accompagnement), l'absence de procédures, un manque de suivi des applications sur les serveurs de la DSI.

On peut réduire ces problèmes à un manque de ressources financières. Ceci étant, sur le terrain, l'expérience ressentie n'est pas l'absence d'argent mais l'absence de personnel dédié et l'absence de moyens informatiques, à deux niveaux, dans les unités de recherche et dans les services communs et centraux.

Ce constat ne minimise pas la réalité de ce que Serres et al. (2017) appellent « le poids des 'écosystèmes' et des pratiques de recherche » dans certaines disciplines : citons la place prépondérante des revues traditionnelles, l'individualisme, la concurrence, ou encore une faible culture du libre accès (p.120). Il existe toujours une « culture papier » revendiquée et assumée, malgré l'utilisation des outils numériques. Cependant, quand il s'agit de bonnes pratiques et de contraintes dues à la législation ou aux programmes de recherche, le verrou majeur n'est pas d'ordre psychologique ou sociologique mais matériel, dû au manque de moyens.

Vers une culture de la donnée

Tous les chercheurs ont affaire aux données, d'une manière ou d'une autre. Savoir collecter, analyser, interpréter, conserver ou communiquer des données fait partie des bonnes pratiques scientifiques. Ce savoir-faire mobilise les valeurs fondamentales de la recherche scientifique, comme l'intégrité, la transparence, l'échange, l'ouverture etc.

L'étude du terrain montre les limites et défaillances des pratiques et outils. Aussi, la politique de la science ouverte crée un nouvel environnement, avec de nouvelles contraintes et injonctions, de la part des agences d'évaluation et de financement, mais aussi de la part des éditeurs, des comités d'éthiques et des délégués à la protection des données (DPD).

Comment améliorer les pratiques ? Comment transformer le savoir-faire en une « culture de la donnée », aussi nommée « data literacy », qui désigne la maîtrise de la façon dont les données sont produites puis exploitées[36] ? L'Université de Bielefeld propose une stratégie de trois piliers (figure 2).

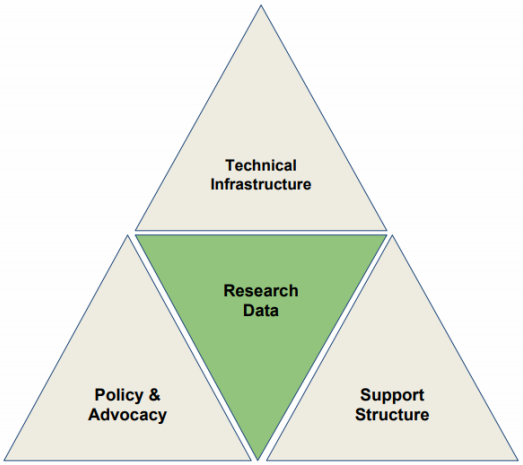

Figure 2. Three Main-Pillars Strategy (Schirrwagen et al. 2018)

L'intérêt de ce modèle est l'intégration des trois dimensions technologie (infrastructure), politique (communication, promotion) et offre de service (accompagnement, soutien). Une particularité de cette stratégie : elle s'articule autour de la création d'une unité interdisciplinaire de formation et de recherche, le *Bielefeld Center for Data Science*[37], et s'oriente actuellement vers un centre de compétence où seront regroupés l'ensemble des services et acteurs sur le campus qui contribuent à la gestion des données de la recherche (curation, droit, dépôt, publication, formation etc.).

Notre propre modèle pour la formation des doctorants développe une approche à trois niveaux, autour des actions de formation, d'une offre de conseil d'assistance personnalisée, et d'un volet technologique (infrastructures) (figure 3).

[36] Deux definitions complémentaires : "Data literacy is the ability to comprehend, create, and communicate data (…) Data-literate individuals have the knowledge, understanding, and skills to connect people to data" SSHRC Knowledge Synthesis project *DataLiteracy.ca* http://dataliteracy.ca/ "(…) a specific skill set and knowledge base, which empowers individuals to transform data into information and into actionable knowledge by enabling them to access, interpret, critically assess, manage, and ethically use data" (Koltay 2016)

[37] BicDas http://www.uni-bielefeld.de/datascience/

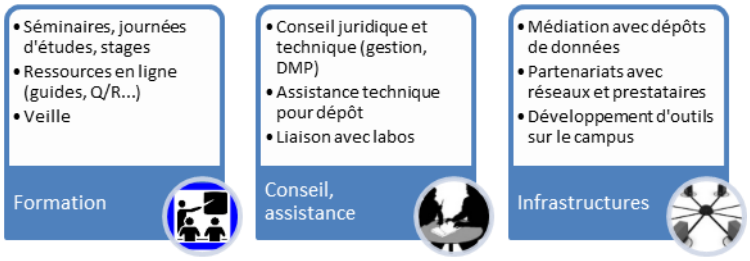

Figure 3. Modèle de Lille (Chaudiron et al. 2015)

Ce type de modèles détermine le cadre d'une stratégie. D'une façon plus concrète, l'étude de Rennes 2 formule six recommandations, déclinées sur trois groupes destinataires, les responsables politiques de la recherche, les responsables des équipes de recherche et les professionnels et chercheurs en IST[38] :

1. Créer un groupe de travail chargé de la réflexion sur une politique des données de recherche ;
2. Développer des services mutualisés pour la sensibilisation et la formation ;
3. Offrir un service juridique spécialisé ;
4. Mieux adapter les espaces internes de stockage des données de recherche aux besoins individuels et collectifs des chercheurs ;
5. Sensibiliser à la description des jeux de données et à la rédaction des Plans de Gestion de Données ;
6. Sensibiliser aux solutions institutionnelles de partage et d'archivage des jeux de données.

A partir de nos propres enquêtes, nos constats et observations, nous formulons dix proposons pour le développement d'une culture de la donnée sur notre campus SHS, avec un pilotage scientifique et en lien avec la politique de l'Open Science du Ministère. Ces propositions sont portées par deux convictions : il faut passer d'une phase de réflexion et de préparation (« groupe de travail ») vers une étape de coordination et de pilotage (« *steering committee* ») ; et plutôt que de développer une stratégie faite de « discours d'injonction, de pression moralisatrice ou culpabilisant sur la question du partage »,[39] qui de toute façon serait condamnée à l'échec, la priorité doit être accordée aux besoins et contraintes auxquels les chercheurs et ingénieurs sont confrontés tous les jours

(1) Mettre en place un pilotage scientifique

Pour développer une culture de données en sciences humaines et sociales, la priorité doit être un pilotage scientifique, par un comité de pilotage et de coordination rattaché à la direction de la recherche, avec des compétences politiques et scientifiques ; ce comité réunirait le Vice-Président Recherche, des représentants des laboratoires et projets scientifiques, des représentants des Directions Recherche (valorisation, ingénierie et management de projets), des représentants du Système d'information et Affaires juridiques, le correspondant du réseau de l'administrateur des données du Ministère, les responsables SSI et RGPD (DPD), le président du comité d'éthique et le directeur du SCD. L'objectif d'un tel comité sera la préparation d'une politique de données à décider par les conseils centraux, et la coordination de sa mise en œuvre.

Par rapport aux enjeux d'une politique de données pilotée par les communautés scientifiques elles-mêmes, la question de l'architecture d'une offre de service paraît secondaire. Trois conditions semblent cependant importantes :

[38] Serres et al. (2017), p.131-136
[39] Serres et al. (2017), p.120

- Une gestion collégiale de cette offre de service par les chercheurs, bibliothécaires, documentalistes et informaticiens.
- Un guichet unique au niveau de la Direction Recherche qui renvoie vers l'offre de services dans les différents services communs et centraux.
- Une offre de service construite sur une approche disciplinaire.

Il faut donner de la valeur à la gouvernance des données pour garantir la qualité et la conformité de la gestion des données de la recherche. Sans aller jusqu'à parler d'un « contrat social entre l'ensemble des utilisateurs à chaque étape de la chaîne des données »[40], il faudra tendre vers l'affichage d'une politique cohérente en matière de données de la recherche, proposant un ensemble d'outils et une offre de services liés aux besoins des chercheurs et des injonctions externes.

(2) Investir d'une manière ciblée

Au terme de leur enquête sur le campus SHS de Rennes 2, Serres et al. (2017) concluent qu'il est impossible de mener une « politique générale, unique et identique pour tous » (p.118). Impossible, en fait, pour deux raisons : non approprié à cause de la richesse des pratiques et la diversité des besoins, et irréaliste du fait des ressources limitées, que ce soit le personnel, les moyens informatiques ou le budget.

Toute politique et stratégie en matière de données doit tenir compte du contexte institutionnel particulier. En ce qui concerne le campus SHS de l'Université de Lille, dans le cadre d'une politique institutionnelle plus large, notre proposition est de limiter l'offre de service « pour tout le monde » à un minimum d'information et de communication en ligne et d'abandonner l'objectif de trouver une solution à tous les problèmes et besoins en matière de données.

Par contre, l'idée est de repenser une stratégie d'acculturation à la donnée à partir de la politique scientifique et de définir des domaines d'action prioritaires. A partir de nos entretiens, mais aussi par rapport à l'offre existante et aux expériences d'autres établissements, voici quelques pistes :

- Focus sur les doctorants, pas seulement pour les plans de gestion mais également pour les espaces de stockage sécurisés et la formation à la gestion des données personnelles et sensibles[41].
- Montage et suivi des projets européens (ou internationaux) et ANR.
- Formation et animation d'un réseau de correspondants RGPD dans les laboratoires.
- Mise en place d'un dispositif de prestation ciblant certains outils et méthodes de recherche, comme des questionnaires, des entretiens qualitatifs, des observations ethnographiques, des mesures physiologiques, des évaluations de traitement etc.

En revanche, nos entretiens tout comme l'expérience de la formation doctorale (séminaire DRTD) laissent penser qu'une approche disciplinaire n'est pas nécessaire ; pas la peine donc de proposer une offre spécifique pour les psychologues, pour les sociologues, pour les archéologues etc.

Investir n'est pas arroser. Concentrer des ressources sur certains domaines d'action prioritaires permet de mettre en œuvre une démarche efficace et efficiente. En plus, si cette démarche est

[40] A. Reyes (2017). `La data literacy ou la "culture de la donnée", le prochain enjeu de nos sociétés'. *Les Echos*, 20 février 2017 https://www.lesechos.fr/idees-debats/cercle/cercle-166504-la-data-literacy-ou-la-culture-de-la-donnee-le-prochain-enjeu-de-nos-societes-2066367.php

[41] Cf. notre livre blanc (Chaudiron et al. 2015) mais aussi la stratégie modèle de l'Université de Wageningen (van Zeeland & Ringersma 2017) ou de l'University College of London (information par Paul Ayris).

accompagnée d'une communication appropriée, on peut compter sur un effet de marketing viral parmi les communautés scientifiques sur le campus.

Ajoutons dans la ligne des préconisations de la LERU la proposition d'intégrer progressivement la gestion des données dans les programmes de Master.

(3) Viser les projets, pas les laboratoires

Souvent, quand il s'agit de projeter une politique d'établissement en matière de données, les laboratoires sont considérés et sollicités comme principal vecteur et courroie de transmission. Or, nos entretiens relativisent fortement cette approche. En fait, le développement d'une culture de la donnée semble plus pragmatique et prometteur si la démarche vise prioritairement les projets de recherche. Ceci pour plusieurs raisons dont :

- Des besoins précis et immédiats.
- Des contraintes imposées par le financement.
- Des obligations légales et réglementaires.
- Une gouvernance plus simple.
- L'expérience des pratiques collaboratives au sein des équipes.

Les enjeux juridiques, éthiques, techniques mais aussi politiques concernent avant tout les projets de recherche, où ils se manifestent non pas comme sujets à débattre mais comme problèmes en attente d'une solution. Malgré un environnement peu comparable, Averkamp et al. (2014) ne disent pas autre chose quand ils observent pour le campus de l'Université d'Iowa que les principaux enjeux de la gestion des données pointent vers des services orientés projets[42].

Actuellement, les contraintes en matière de stockage et d'archivage, mais également de diffusion, sont particulièrement fortes dans les projets européens (programme H2020), mais il faut anticiper l'adoption d'une politique similaire par les programmes de l'ANR, dans le cadre de la politique de la science ouverte.

Il serait donc raisonnable de focaliser la politique d'établissement ou de site sur les projets d'envergure, européens, internationaux et nationaux. En tête, les projets H2020 et ANR, mais il ne faudrait pas négliger d'autres programmes de financement, comme les subventions du fonds régional FEDER, par exemple.

Plusieurs actions faciliteraient la mise en œuvre d'une telle approche :

- Centraliser les dossiers de recherche au niveau du RSSI.
- Cibler le conseil et l'assistance du RSSI sur les projets, pour sensibiliser et adopter de bonnes pratiques etc.
- Côté Direction Recherche, proposer un kit de montage qui tient compte de la dimension « données », à destination des chercheurs et des professionnels impliqués (bibliothécaires, documentalistes, ingénieurs), adapté par exemple du service DoRANum[43] du CNRS ou du cours en ligne de *Research Data Netherlands*[44].
- Intégrer systématiquement le plan de gestion dans le cahier des charges lors du montage des projets.

[42] "The main challenges point to a need for project or discipline-specific services."
[43] http://doranum.fr/
[44] *Essentials 4 Data Support* par rdnl http://datasupport.researchdata.nl/en/

(4) Utiliser les plans de gestion comme levier

Pour développer une culture de la donnée et mettre en œuvre de bonnes pratiques de gestion, les plans de gestion représentent probablement le meilleur levier. Ils sont devenus obligatoires pour les projets européens du programme H2020, ils figurent parmi les actions prioritaires du Plan d'action national 2018-2020, et l'ANR a annoncé vouloir s'aligner sur la Commission Européenne à partir des appels à projets pour 2019.

Dans la mesure où ils décrivent l'ensemble du cycle de vie de la gestion des données, de la collecte au traitement et à la génération de nouvelles données, les plans de gestion des données constituent un élément clé de la bonne pratique en matière de données[45]. Concrètement, un bon plan devrait informer sur au moins cinq aspects :

- le traitement des données de recherche pendant et après la fin du projet,
- quelles données seront collectées, traitées et/ou générées,
- quelle méthodologie et quelles normes seront appliquées,
- si les données seront partagées et/ou diffusées en open access,
- comment les données seront indexées et conservées (y compris après la fin du projet).

Il y a quelques années, les deux agences américaines *National Science Foundation* et *National Institutes of Health* ont fait du plan de gestion et du partage des données une condition de financement de projet. Averkamp et ses collègues de l'University of Iowa ont décrit comment cette politique de données a changé les pratiques, attitudes et services sur le campus[46].

Notre proposition est d'utiliser ces plans comme levier pour la mise en place d'une offre de service et pour le développement d'une culture de la donnée. L'idée est de s'appuyer sur une contrainte forte externe, liée aux projets les plus prestigieux et les mieux dotés, aujourd'hui H2020, demain ANR.

Une telle approche pourrait adopter le cadre référentiel du groupe de travail de Science Europe pour établir des protocoles standards pour la gestion des données de la recherche, faisant le lien entre des pratiques communautaires et les besoins des infrastructures (principe FAIR)[47].

Cette approche devrait avant tout établir le lien entre les plans de gestion et des protocoles éthiques, au niveau de l'offre de service, des procédures et du suivi. Tous les plans contiennent un volet éthique, et tous les protocoles éthiques contiennent des informations sur la nature et le traitement des données. Ainsi, l'exemple de la trame de protocole de l'Université de Lille de mai 2018 destiné à des études en sciences comportementales prévoit un chapitre 4 « Traitement des données » avec trois sections :

4.1. Gestion des données

Préciser où seront stockés les données nominatives (la liste d'identification code-nom participant), les consentements et lettres d'information.

---

[45] European Commission (2016)

[46] "Awareness of data management has been increasing as NSF has begun to require data management plans and NIH is enforcing data sharing plans. Many participants required to write plans indicated that they receive help for data management planning from colleagues or support services at the department or college level." (Averkamp et al. 2014)

[47] « An RDM protocol would contain the usual elements of a DMP. It would pay special attention to standards and guidelines for data management that are relevant for a specific field or research community that shares similar data collection and processing methods. It would be a public document that could be properly referenced and should be considered a template that is already mostly filled in, possibly offering alternatives from which a researcher can choose, depending on the particularities of their research project." (Doorn et al. 2018)

Préciser où seront stockés les données non nominatives (réponses aux questionnaires, données socio démo, résultats à l'expérience).

Dire qui s'occupe de la saisie des données, qui a accès aux données...

4.2. Analyse des données

Noter les critères d'évaluation principaux et secondaires.

Préciser quel type de traitement statistique sera effectué, selon le type de données et les hypothèses.

4.3. Finalité de l'étude

Préciser les résultats attendus et dégager des perspectives.

Le protocole souligne l'importance de ce passage car il sera visé par le responsable DPD (Délégué à la Protection des Données) qui participe aux comités d'éthique.

Dans le cycle d'un projet de recherche, le plan de gestion et le protocole éthique appartiennent à la phase de préparation, et il fait sens, du point de vue de chercheur, de les lier d'une manière ou d'une autre, afin de réduire la charge de travail, d'optimiser et rationnaliser le dispositif sur le campus.

Il y a une autre raison pour lier les deux instruments : l'obligation des projets H2020 de produire un plan de gestion à mi-parcours et un autre à la fin. Cette obligation pourrait être une opportunité pour améliorer le suivi de projets au niveau des données, avec notamment des audits ponctuels sur les deux aspects.

(5) Apporter des réponses aux contraintes de sécurité

Une politique institutionnelle en matière de données de la recherche qui se limiterait à l'injonction à la conservation et au partage des données, sans apporter des réponses concrètes aux questions de la sécurité des données et des systèmes manquerait de crédibilité. La sécurité est la première préoccupation des chercheurs, et l'établissement doit proposer un environnement de travail à la hauteur des enjeux et des obligations réglementaires, sur ses propres serveurs et/ou « in the cloud », avec des partenaires publics ou privés. Quelques éléments, à partir des entretiens :

- Assurer une protection des données contre le piratage et d'autres risques (incendie, vol, crash…), pendant la durée du projet.
- Proposer un système de sauvegarde et de récupération des données pendant et après le projet.
- Ne pas séparer la protection des données de la sécurité d'autres applications des projets et structures de recherche (sites web, serveurs revue etc.).
- Etablir une analyse des risques et sinistres informatiques, avec des cas de figure (incidents), scénarios etc., à destination des chercheurs.
- Réaliser une analyse des risques liés à un projet, en amont (phase de préparation), d'une manière collégiale.
- Faire de temps en temps un audit sécurité au niveau d'un projet ou d'un laboratoire, en commençant par les endroits moins sensibles pour « ne pas faire peur aux gens ».

Quatre autres remarques pour finir : La fusion pourrait être l'opportunité pour rééquilibrer les ressources en matière de sécurité informatique et de personnel SSI entre les campus STM et SHS. Même si les données sensibles sont prioritaires, il faut tenir compte que tout le monde risque de perdre de données. Une politique d'établissement devrait inclure d'une manière ou d'une autre les initiatives et dispositifs de données adoptés dans les CHU ; en effet certains projets SHS sont menés

dans les hôpitaux et la santé est l'un des domaines prioritaires de la politique de recherche de la région et de l'Université de Lille. Et dernière remarque, les enquêtes sur la gestion des données en SHS décrivent des pratiques centrées sur l'ordinateur personnel (privé et/ou professionnel). Une démarche sécurité pourrait changer la situation, en mettant le cloud au cœur de la gestion et de la protection des données (cf. figure 4).

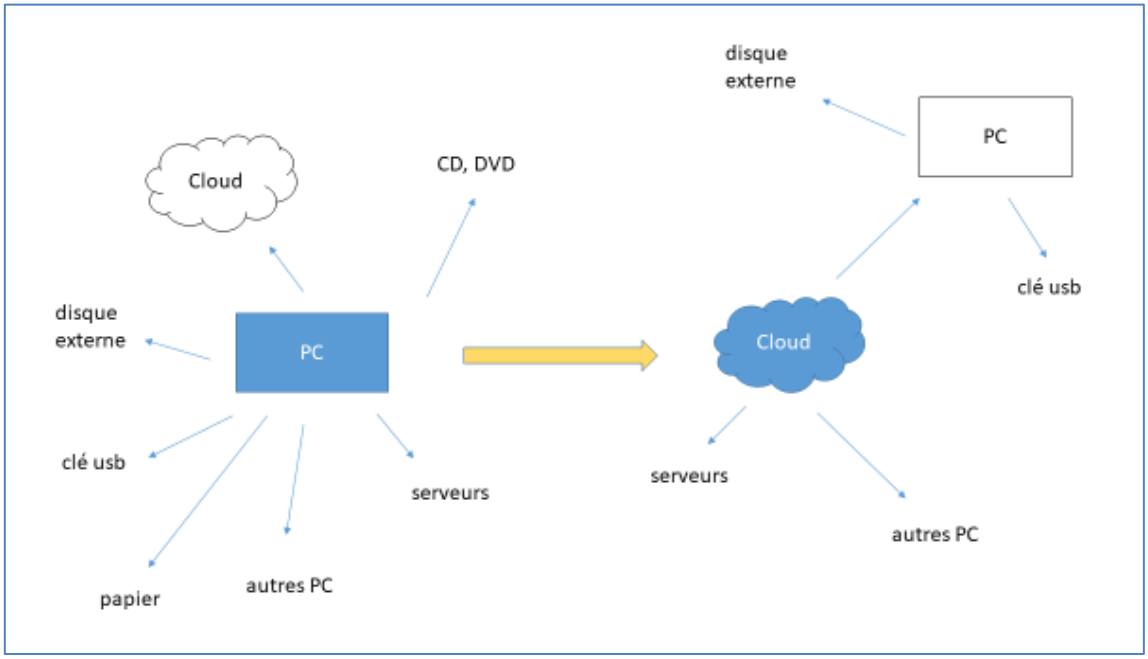

Figure 4. De l'ordinateur personnel vers le cloud

(6) Apporter des réponses aux besoins de communication

La communication sécurisée des données au sein d'une équipe projet et avec d'autres établissements et collègues est l'autre préoccupation des chercheurs. Une politique de données doit faire de ce besoin une deuxième priorité.

De nouveau, il n'y a sans doute pas une seule solution, et la solution ne se trouve pas nécessairement sur le campus et au sein de l'établissement. Néanmoins, dans le cadre d'un dispositif de données, l'option la plus réaliste serait aujourd'hui un serveur de stockage temporaire avec des fonctionnalités de partage, semblable à Dataverse, avec des instances communautaires.

(7) Apporter des réponses aux besoins de curation

Pour commencer, aucun de nos interlocuteurs n'a utilisé le mot « curation » et très peu ont exprimé le besoin d'un conseil ou de l'assistance pour la description et la gestion des jeux de données. Cependant, le besoin est réel, imposé par la politique de données nationale et européenne en faveur d'une science ouverte et de l'interopérabilité des infrastructures.

Il s'agit en priorité de trois domaines, en lien avec les principes FAIR :

Contribuer à la normalisation des métadonnées. Ceci aussi bien pour des formats génériques (Dublin Core, Datacite Metadata Schema…) que pour des formats spécifiques, liés à des domaines ou outils particuliers (Data Documentation Initiative…).

Contribuer à l'utilisation (attribution) d'identifiants uniques. Pour les données, des DOI (coopération avec l'initiative Datacite), handle ou des identifiants plus spécifiques ; d'autres identifiants pour les auteurs (ORCID etc.) ou les institutions.

Contribuer à la création de liens avec les publications associées. Avant tout, créer des liens avec les métadonnées dans l'archive ouverte institutionnelle, éventuellement aussi sur d'autres plateformes.

Comme pour les publications, il s'agit d'activités et de compétences des métiers de l'information, au sein du SCD, des bibliothèques associées et des services de documentation dans les laboratoires, ce qui nécessite une coopération entre bibliothécaires, documentalistes, chercheurs et, le cas échéant, ingénieurs de données.

Une telle offre de service doit rester flexible et s'adapter aux infrastructures, politiques et pratiques des chercheurs, en proposant aussi bien de la médiation, de la formation et de l'assistance, mais aussi de la « sous-traitance », via la prise en charge de la curation par les professionnels. Aussi, une partie de la curation demande une expertise de spécialistes d'un domaine ou d'un équipement, qui se trouve davantage dans les laboratoires que dans les services communs.

(8) Proposer plusieurs solutions pour la conservation des données

Nous avons constaté une large gamme de pratiques et besoins en matière de conservation. Aussi, le terme même de conservation englobe, du point de vue fonctionnel mais aussi informatique, plusieurs dispositifs. Il suffit de rappeler les dispositifs de la conservation à court terme (archives vivantes), pour la durée d'un projet ou de sa valorisation (publications etc.), les entrepôts et plateformes pour la conservation à plus long terme (archives définitives) pour une capitalisation, une réutilisation ou bien, imposée par la loi comme pour la recherche dans le domaine de la santé (15 ans), et enfin les infrastructures pour une conservation à long terme sans diffusion (« dark archive », telle le CINES[48]).

Proposer des solutions ne veut pas dire, développer toute une gamme de dispositifs sur le campus ; ce serait irréaliste. Mais il faudrait pouvoir proposer des espaces appropriées (en termes de volumétrie, sécurité, accessibilité) pour les besoins des chercheurs, soit sur les serveurs de l'université, soit en partenariat avec des prestataires externes.

Dans l'environnement et à l'instar de l'archive ouverte institutionnelle se posera sans doute très vite la question de la faisabilité et de l'intérêt d'une solution locale pour la conservation d'une partie des données sur un serveur institutionnel, avec des systèmes comme Dataverse ou Invenio, comme solution de conservation par défaut.

Pour le reste, il s'agira de conseiller, d'orienter, le cas échéant de faire le lien avec des infrastructures nationales (par exemple Huma-Num[49]), internationales et/ou disciplinaires.

Ajoutons l'intérêt d'espaces intranet pour la documentation des protocoles éthiques et des plans de gestion, un projet en cours à la DSI.

(9) Institutionnaliser le lien avec la TGIR Huma-Num

La TGIR Huma-Num[50] a été mise en œuvre par le Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation en 2013. Elle est portée par le CNRS, l'Université d'Aix-Marseille et le

[48] Centre Informatique National de l'Enseignement Supérieur https://www.cines.fr/
[49] cf. plus loin, ou avec un éventuel service d'accueil et de diffusion pour des données simples préconisé par le plan national du Ministère
[50] Très grande infrastructure de recherche des humanités numériques Huma-Num http://www.huma-num.fr/

Campus Condorcet. Seule grande infrastructure de recherche dédiée aux lettres, sciences humaines et sociales et aux humanités numériques, Huma-Num propose des services numériques pour les programmes de recherche et anime un réseau de consortiums sur les humanités numériques[51].

Huma-Num propose une gamme de services originale pour la gestion des données. Un partenariat avec le CINES assure la conservation à long terme des données déposées sur la plateforme NAKALA[52]. Cette plateforme s'inscrit dans un ensemble de services pour faciliter l'accès, le signalement, la conservation et l'archivage à long terme des données de la recherche en SHS.

La plateforme Huma-Num s'adresse avant tout aux équipes et laboratoires de recherche. Certains laboratoires, dont SCALab, ont déjà pris contact avec Huma-Num pour trouver une solution à l'archivage de leurs données de la recherche. Afin de promouvoir le dispositif, de coordonner et de faciliter les contacts avec les SHS de Lille, nous suggérons la désignation d'un correspondant local TGIR Huma-Num au sein de la bibliothèque universitaire SHS, à l'instar de l'Université de Nice Sophia Antipolis, où la MSH Sud-Est et le SCD ont mis en place un tel correspondant en 2017[53]. Cette nomination pourrait s'appuyer sur l'expérience du projet *D4Humanities* qui a fait la connexion avec NAKALA pour les jeux de données des doctorants de l'Ecole Doctorale SHS.

Un tel partenariat ouvrirait trois autres perspectives, dans le domaine des données :

- Renforcer les liens entre SCD et MESHS.
- Contribuer au développement de l'offre de service du TGIR.
- Développer la présence de l'Université de Lille dans les infrastructures européennes DARIAH[54] et CLARIN[55] dans lesquelles Huma-Num représente la France.

(10) Soutenir les bonnes pratiques

La dernière proposition est de valoriser les bonnes pratiques sur le campus, à titre d'exemple et de modèle, pour la communication, pour la formation, aussi pour la promotion et le marketing de nouveaux services et outils. Cette valorisation pourrait également permettre de faire émerger certaines initiatives dans le domaine des données de la recherche, par rapport à la création d'un prix récompensant les équipes et projets exemplaires annoncé par le plan d'action du Ministère.

Aux Pays Bas, l'Université de Wageningen a développé sa stratégie en matière de données à partir de l'analyse et de la valorisation des bonnes pratiques, notamment pour crédibiliser et illustrer leurs actions de communication, d'information et de formation[56].

Une telle stratégie impliquera plusieurs étapes :

- Identifier les cas de bonnes pratiques, par discipline, laboratoire, équipement, méthodologie ou type de projet.

[51] Cf. le rapport d'activité 2013-2015 sur HAL https://halshs.archives-ouvertes.fr/halshs-01390938/
[52] https://www.nakala.fr/
[53] Cf. http://mshs.unice.fr/?cat=38
[54] Digital Research Infrastructure for the Arts and Humanities https://www.dariah.eu/
[55] European Research Infrastructure for Language Resources and Technology https://www.clarin.eu/
[56] Cf. von Zeeland & Ringersma (2017) : "Credibility of the guidelines is likely enhanced due to them being built on best-practices in the organisation. Use cases that store their data on safe solutions, archive their data in trusted repositories following the FAIR principles, and/or consistently register their datasets, show other research groups how the guidelines can be followed in practice. In fact, if these best-practice use cases do not object, their data management practices will be used in communicating the new policy, serving as exemplars to other researchers."

- Décrire ces cas dans leur contexte, avec leurs facteurs-clés de succès, leurs retombés etc. (« story-telling »).
- Créer des vitrines virtuelles pour les rendre visibles au plus grand nombre (« showcases »).
- Construire du matériel de communication (vidéos, sites, plaquettes…) et de formation (recommandations, guides, procédures, modélisation, « Kit données ») à partir des exemples de bonnes pratiques.

Cette valorisation pourrait cibler, du moins au début, certains domaines et cas de figures prioritaires, comme les recherches en santé, les projets européens, les enquêtes, les enregistrements vidéo, le stockage sécurisé, le traitement des données dans le protocole éthique, les outils de communication (transfert de fichier) ou l'utilisation d'un entrepôt comme Zenodo.

Conclusion

La science ouverte figure désormais parmi les priorités de l'Etat français. Le plan national du Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation poursuit l'objectif que les données produites par la recherche publique soient progressivement structurées en conformité avec les principes FAIR, qu'elles soient préservées et, dans la mesure du possible, qu'elles soient ouvertes.

Comment mettre en œuvre l'écosystème de la science ouverte sur le terrain d'un campus en sciences humaines et sociales ? Comment développer une culture de la donnée ? A partir de 51 entretiens avec des enseignants-chercheurs et ITRF du campus SHS (Pont de Bois) de l'Université de Lille, notre étude propose un cadre global. Elle a été réalisée dans le cadre du projet structurant *D4Humanities*, avec un financement de la MESHS et du Conseil Régional Hauts-de-France, et elle fait suite à des travaux de recherche menés depuis 2013 par le laboratoire GERiiCO.

Son objectif principal : (re)mettre les enseignants-chercheurs au cœur d'une politique de données sur le campus, avec leurs besoins, priorités et interrogations. Après une analyse des pratiques et besoins, et après avoir identifié des opportunités et verrous, l'étude fait dix recommandations dont l'essentiel tient en trois points :

1. Mettre en place un pilotage scientifique, pour une coordination des actions et services de l'ensemble des acteurs (services centraux et communs, structures de recherche, composantes etc.).
2. Concentrer la politique sur certaines actions ciblées, en tenant compte des priorités des chercheurs et en mettant l'accent sur les projets de recherche (H2020, ANR etc.).
3. Positionner la démarche clairement au sein des infrastructures nationales et européennes en SHS, en particulier par une institutionnalisation des liens avec la TGIR Huma-Num.

En revanche, il faut éviter tout discours d'injonction idéologique sur la question de l'ouverture des données, tout comme il faut éviter l'éparpillement des efforts et ressources. Il n'est pas possible de donner une réponse à toutes les demandes, d'autant qu'une partie des solutions se trouvent à l'extérieur du campus, dans les projets et communautés de recherche, dans les infrastructures et services au niveau national, et dans les réseaux internationaux. La politique à mener devrait appliquer des principes de subsidiarité et de complémentarité, ce qui implique une très bonne connaissance du terrain de la recherche et des dispositifs de données.

Deux autres actions devraient accompagner une telle politique de campus :

1. La création d'une bibliothèque de référence sur les données et la science ouverte, regroupant des rapports, études, standards, manuels, recommandations, ouvrages de références etc.
2. La mise en place d'un projet scientifique multidisciplinaire pour analyser les besoins et pratiques, pour évaluer les politiques, dispositifs et modèles économiques, et pour assurer le suivi de la mise en œuvre d'un écosystème de la science ouverte.

Le développement d'une culture de la donnée ne se limitera pas à la mise en place de nouveaux services et au changement de pratiques, mais aura besoin d'une analyse critique, d'une prise de recul, d'une compréhension des enjeux dans la meilleure tradition universitaire.

Références

F. André (2015). `Déluge des données de la recherche ?'. In L. Calderan, P. Laurent, H. Lowinger, & J. Millet (eds.), *Big data : nouvelles partitions de l'information. Actes du Séminaire IST Inria, octobre 2014*, pp. 77-95. De Boeck; ADBS, Louvain-la-Neuve.

S. Averkamp, et al. (2014). `Data Management at the University of Iowa: A University Libraries Report on Campus Research Data Needs'. University of Iowa. http://ir.uiowa.edu/lib_pubs/153/

B. Bauer, et al. (2015). `*Researchers and Their Data. Results of an Austrian Survey - Report 2015*'. e-infrastructures austria, Vienna. https://e-infrastructures.at/de

C. L. Borgman (2016). *Big data, little data, no data: scholarship in the networked world*. The MIT Press, Cambridge MA.

S. Chaudiron, et al. (2015). *Livre blanc sur les données de la recherche dans les thèses de doctorat*. Université de Lille 3, Villeneuve d'Ascq. https://hal.archives-ouvertes.fr/GERIICO/hal-01192930

COMETS (2015). *Les enjeux éthiques du partage des données scientifiques*. Comité d'éthique du CNRS, Paris. http://www.cnrs.fr/comets/spip.php?article123

DCC (2018). *Data management. Learning from the innovators: 2 leading funders compared*. SPARC Europe, Apeldoorn. https://sparceurope.org/new-case-study-examines-two-winning-open-data-policy-drivers/

F. Debos (2017). `Open data et culture de la donnée : le cas OpeNRJ'. *Revue française des sciences de l'information et de la communication* (10). https://journals.openedition.org/rfsic/2639

P. Doorn (ed.) (2018). *Science Europe Guidance Document. Presenting a Framework for Discipline-specific Research Data Management*. Science Europe Working Group on Research Data, Brussels. https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf

Etalab (2018). `*Pour une action publique transparente et collaborative : plan d'action national pour la France 2018-2020*'. Secrétariat d'Etat chargé de la Réforme de l'Etat et de la Simplification, Paris. https://www.etalab.gouv.fr/wp-content/uploads/2018/04/PlanOGP-FR-2018-2020-VF-FR.pdf

European Commission (2016). `*H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020.Version 3.0*'. European Commission Directorate-General for Research & Innovation, Brussels. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

T. E. Hardwicke, et al. (2018). `Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition'. *BITSS*. http://doi.org/10.17605/OSF.IO/39CFB

P. B. Heidorn (2008). `Shedding Light on the Dark Data in the Long Tail of Science'. *Library Trends* 57(2):280-299. https://www.ideals.illinois.edu/bitstream/handle/2142/10672/heidorn.pdf?sequence=2

B. Jacquemin, et al. (2013). `Ouvrir les données de la recherche pour la veille scientifique. Le cas des thèses électroniques'. In *VSST'2013, Nancy, 23-25 octobre 2013*. http://archivesic.ccsd.cnrs.fr/sic_01070495/fr/

J. Klump (2017). `Data as Social Capital and the Gift Culture in Research'. *Data Science Journal* 16(14):1-8. https://datascience.codata.org/articles/10.5334/dsj-2017-014/

T. Koltay (2016). `Data literacy for researchers and data librarians'. *Journal of Librarianship and Information Science* 49(1):3-14. http://journals.sagepub.com/doi/10.1177/0961000615616450

T. Kuipers & J. van der Hoeven (2009). `*Insight into digital preservation of research output in Europe. Survey report*'. PARSE insight, n/a. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

LERU (2018). `*Open science and its role in universities: A roadmap for cultural change*'. League of European Research Universities, Leuven, Belgium. https://www.leru.org/publications/open-science-and-its-role-in-universities-a-roadmap-for-cultural-change

LIBER working group on E-Science / Research Data Management (2012). `*Ten recommendations for libraries to get started with research data management*'. LIBER Association of European Research Libraries, The Hague. http://libereurope.eu/wp-content/uploads/The%20research%20data%20group%202012%20v7%20final.pdf

MESRI (2018). `*Plan national pour la science ouverte*'. Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, Paris. http://m.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html

B. Michel & V. Chekib (2015). `*Les pratiques de recherche documentaire, de publication et de diffusion scientifique des productions de la recherche à l'Université Paris-Sud : questionnaire à destination des chercheurs, enseignants-chercheurs et doctorants*'. Université Paris-Sud. https://hal.archives-ouvertes.fr/hal-01292693/

P. Naegelen (2015). `Données de la recherche : quel positionnement et quels rôles pour les bibliothèques ?'. In *Données en partage : enjeux et acteurs des données de la recherche. URFIST Toulouse, 15 juin 2015*. http://fr.slideshare.net/pierrenaegelen/donnes-de-la-recherche-quel-positionnement-et-quels-rles-pour-les-bibliothques

H. Prost & J. Schöpfel (2015). `*Les données de la recherche en SHS. Une enquête à l'Université de Lille 3. Rapport final*'. Université de Lille 3, Villeneuve d'Ascq. http://hal.univ-lille3.fr/hal-01198379v1

G. Pryor, et al. (dir.) (2014). *Delivering research data management services: fundamentals of good practice*. Facet, London.

A. Rege (2015). `Retour sur l'enquête sur les pratiques de publication scientifique et de production de données de la recherche'. In: *Université de Strasbourg, Projet AOC, COPIL* 27 mars 2015.

S. Reilly, et al. (2011). `*Report on Integration of Data and Publications*'. ODE Opportunities for Data Exchange, The Hague. http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf

R. Reznik-Zellen, et al. (2012). `Tiers of Research Data Support Services'. *Journal of eScience Librarianship* 1(1):27-35. http://escholarship.umassmed.edu/jeslib/vol1/iss1/5/

J. Schirrwagen, et al. (2018). `Expanding the research data management service portfolio at Bielefeld University according to the three-pillar principle towards data FAIRness'. In Göttingen-CODATA RDM Symposium 2018, 2018-03-18 – 2018-03-20, Göttingen. https://pub.uni-bielefeld.de/publication/2919659

J. Schöpfel, et al. (2017). `« Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse'. In *Atelier VADOR : Valorisation et Analyse des Données de la Recherche, INFORSID 2017, 31 mai 2017 Toulouse*. http://hal.univ-lille3.fr/hal-01530937v1

J. Schöpfel, et al. (2018). `Research Data Management in the French National Research Center (CNRS)'. *Data Technologies and Applications* 52(2):248-265. https://www.emeraldinsight.com/doi/abs/10.1108/DTA-01-2017-0005

A. Serres, et al. (2017). `*Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2*'. Université Rennes 2. https://hal.archives-ouvertes.fr/hal-01635186

E. Simukovic, et al. (2013). `Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin'. *Humboldt-Universität zu Berlin, Zentraleinrichtung Computer- und Medienservice (Rechenzentrum)*, Berlin. http://edoc.hu-berlin.de/docviews/abstract.php?id=40341

E. Simukovic, et al. (2014). `Unveiling Research Data Stocks: A Case of Humboldt-Universität zu Berlin'. In *iConference, 4-7 March 2014, Berlin*, pp. 742-748. http://hdl.handle.net/2142/47259

M. D. Wilkinson, et al. (2016). `The FAIR Guiding Principles for scientific data management and stewardship'. *Scientific Data* 3:sdata201618+. https://www.nature.com/articles/sdata201618

H. van Zeeland & J. Ringersma (2017). `The development of a research data policy at Wageningen University & Research: best practices as a framework'. *LIBER Quaterly* 27(1):153-170. https://www.liberquarterly.eu/article/10.18352/lq.10215/

Annexe 1 Trame des entretiens

1. Identifiant de l'enquête 2015
2. Quel est votre statut (PR, MCF, doctorant...) ?
3. A quel(s) laboratoire(s) appartenez-vous ?
4. Quel est le numéro de votre section CNU ?
5. Quelles sont vos thématiques de recherche ?
6. Participez-vous (ou avez-vous participé) à un projet ANR ?
   - Oui
   - Non
   - J'en ai l'intention
7. Participez-vous (ou avez-vous participé) à un projet européen (H2020 ou autre) ?
   - Oui
   - Non
8. Connaissez-vous les consignes du programme H2020 pour la gestion des données de la recherche ?
   - Oui
   - Non
9. Connaissez-vous d'autres consignes pour la gestion des données de la recherche ?
   - Oui
   - Non
10. Quel(s) type(s) de données recueillez ou utilisez-vous ? Constituez-vous des jeux de données ?
    - Document Texte
    - Enquête et entretiens
    - Observation
    - Expériences
    - Archives
    - Statistique

- Photos d'objets
- Documents audio et vidéo
- Autre

11. Sous quel(s) format(s) sont enregistrés les fichiers contenant vos données ?

12. Quel(s) type(s) de données produisez-vous ? De quel(s) format(s) sont-elles ?

- Texte
- Tableaux
- Bases de données
- Visualisations ou modèles multidimensionnelles
- Enregistrement audio
- Photos
- Cartes et plans
- Programmes et applications
- Autre

13. Travaillez-vous avec des données personnelles (au sens de la CNIL) ?

- Oui
- Non

14. Travaillez-vous avec des données confidentielles (partenariats avec entreprises...) ?

- Oui
- Non

15. Avez-vous déjà soumis un protocole de recherche au comité d'éthique de l'Université de Lille 3 (ou d'un autre établissement, CNRS etc.) ?

- Oui
- Non

16. Dans votre discipline, y a-t-il des outils ou guides pour gérer les données ? Quels sont-ils ?

- Oui
- Non

17. Etes-vous satisfait de ces outils ? Si non, que vous manque-t-il ?

    - Oui
    - Non

18. Avez-vous déjà rédigé un plan de gestion des données ?

    - Oui
    - Non

19. Si oui, dans quelles conditions l'avez-vous rédigé ?

    - Par le comité d'éthique
    - Avec le responsable de la sécurité informatique
    - Autre

20. Attribuez-vous des identifiants (numéros, codes…) à vos données ? Suivez-vous des règles particulières pour nommer vos données ?

    - Oui
    - Non

21. Comment décrivez-vous les données ? Avec quels éléments, quelles métadonnées ?

    - Intitulé du fichier
    - Auteur
    - Domaine, thématique
    - Droits (protection, confidentialité)
    - Méthodologie (outils)
    - Projet (nom, financement, date)
    - Autre

22. Comment classez-vous (/organisez-vous) vos données ?

23. Pour cette description, suivez-vous une norme ou recommandation ? Si oui, précisez laquelle (Norme ISO, AFNOR ou norme standard dans la discipline)

    - Oui
    - Non

24. Etes-vous censé(e) publier vos données ?

    - Oui
    - Non

25. Partagez-vous vos données avec d'autres ? Si oui, précisez avec qui ?

    - Oui
    - Non

26. Comment diffusez-vous vos données lorsque le projet de recherche est en cours ?

    - Site web personnel
    - Archive thématique
    - A la demande
    - Autre

27. Comment diffusez-vous vos données une fois le projet de recherche terminé ?

    - Site web personnel
    - Archive thématique
    - A la demande
    - Autre

28. Avez-vous déjà déposé vos données sur un serveur en ligne ? Si oui, précisez lequel

    - Oui
    - Non
    - Pas encore, j'ai l'intention de la faire

29. Si oui, à part vous qui pouvait/peut accéder à ces données ?

    - Personne
    - Equipe projet
    - Tout le monde

30. Comment s'opère l'accès à vos données ?

    - Accès libre
    - A la demande
    - Avec autorisation ou enregistrement
    - Autre

31. Utilisez-vous des licences de réutilisation pour le partage de vos données (licence ouverte, Creative Commons etc.) ?
    - Oui
    - Non
32. Bénéficiez-vous de conseils dans cette démarche ?
    - Oui
    - Non
33. Où stockez-vous vos données? Existe-t-il une différence dans la solution de stockage utilisée pour les données brutes, les données analysées et les données de sources externes ?
34. Pouvez-vous donner une estimation de l'espace de stockage nécessaire pour vos données durant vos recherches ?
35. Pouvez-vous donner une estimation de l'espace nécessaire pour le stockage de vos données après la fin de vos recherches ?
36. Sur quel(s) support(s) effectuez-vous ce stockage ?
    - Disque dure
    - Serveur en ligne
    - Autre
37. Pouvez-vous nous exposer étape par étape les différentes procédures que vous suivez pour la gestion de vos données - qui y participe, avec quels logiciels ou outils, quels problèmes ?
    - Collecter :
    - Analyser :
    - Stocker :
38. Quel type de conseil (formation, logiciels/outils, etc.) vous aiderait à simplifier ces tâches ?
39. De quels services auriez-vous besoin pour la gestion de vos données ?
40. Avez-vous d'autres questions liées à la gestion des données ?

Annexe 2 Liste des interlocuteurs

Nom	Prénom	Laboratoire	Fonction/statut
Antoine	Pascal	SCALab	Enseignant-chercheur
Azouz	Kaouther	GERiiCO	Doctorant
Benoist	Stéphane	HALMA	Direction laboratoire
Benoit	Martine	CECILLE	Direction MESHS
Bobas	Constantin	CECILLE	Direction laboratoire
Brasseur	Pierre	CeRIES	Doctorant
Brunellière	Angèle	SCALab	Enseignant-chercheur
Canut	Emmanuelle	STL	Enseignant-chercheur
Cassette	Eric	DSI	RSSI
Castellani	Marie-Madeleine	ALITHILA	Direction laboratoire
Chaudiron	Stéphane	GERiiCO	Direction laboratoire
Christophe	Véronique	SCALab	Enseignant-chercheur
Clerc	Jérôme	PSITEC	Enseignant-chercheur
Corveleyn	Xavier	SCALab	Enseignant-chercheur
Courbois	Yannick	PSITEC	Enseignant-chercheur
Dabo	Sophie	LEM	Enseignant-chercheur
David	Hélène	ALITHILA	Doctorant
De la Broise	Patrice	GERiiCO	Direction laboratoire
Delevoye	Yvonne	SCALab	Comité d'éthique
Dereymaeker	Nathalie	IRHiS	Doctorant
Desbiens	Agnès	PSITEC	Enseignant-chercheur
Deville	Julie	CIREL	Enseignant-chercheur
Dutoit	Thomas	CECILLE	Direction laboratoire
Fraisse	Amel	GERiiCO	Enseignant-chercheur
Gamelin	Thomas	HALMA	Postdoc
Gawin	Geoffroy	GERiiCO	Doctorant
Grabar	Natalia	STL	Chercheur
Jenn	Ronald	CECILLE	Enseignant-chercheur
Jozefowiez	Jérémie	SCALab	Enseignant-chercheur
Kozlowski	Lise	Direction recherche	Gestion juridique
Ladrouz	Mohamed	DGS	CIL
Lassus	Marie-Pierre	CECILLE	Enseignant-chercheur
Lehoërff	Anne	HALMA	Enseignant-chercheur
Lelorain	Sophie	SCALab	Enseignant-chercheur
Leoni	Véronique	PSITEC	Enseignant-chercheur
Lesenne	Sabine	GERiiCO	Postdoc
Maltet	Zoé	GERiiCO	Doctorant

Mc Intosh	Fiona	ALITHILA	Direction laboratoire
Messing	Sabrina	ALITHILA	Doctorant
Micheau	Béatrice	GERiiCO	Enseignant-chercheur
Planckeel	Charlotte	HALMA	Doctorant
Proust	Sophie	CEAC	Enseignant-chercheur
Robinne	Christophe	IRHiS	Doctorant
Roger	Clémence	SCALab	Enseignant-chercheur
Sampson	Marie-Pierre	HALMA	Gestion laboratoire
Sève	Laurianne	HALMA	Direction laboratoire
Sparrow	Laurent	SCALab	Enseignant-chercheur
Tessier	Jean-Luc	DGS	CIL / DPDP
Timbert	Arnaud	IRHiS	Enseignant-chercheur
Truel	Myriam	CECILLE	Doctorant
Zieger	Karl	ALITHILA	Direction laboratoire